

Linear Context-Free Tree Languages and Inverse Homomorphisms

Johannes Osterholzer Toni Dietze Luisa Herrmann

Faculty of Computer Science, Technische Universität Dresden

October 20, 2015

We prove that the class of linear context-free tree languages is not closed under inverse linear tree homomorphisms. The proof is by contradiction: we encode Dyck words into a context-free tree language and prove that its preimage under a certain linear tree homomorphism cannot be generated by any context-free tree grammar. A positive result can still be obtained: the linear monadic context-free tree languages *are* closed under inverse linear tree homomorphisms.

1 Introduction

Context-free tree grammars (cftg), introduced by Rounds [23], generalize the concept of context-free rewriting to the realm of tree languages. A sentential form of a cftg is a tree labeled by terminal and nonterminal symbols. In contrast to a regular tree grammar, any node of a sentential form may be labeled by a nonterminal, not just a leaf node. In general, cftg can *copy* parts of a sentential form in the application of a production, and a lot of their complexity in comparison to regular tree grammars is due to the interplay between copying and nondeterminism (cf. e.g. [7]).

While Rounds viewed context-free tree grammars as a promising model for mathematical linguistics, the main motivation for studying cftg in the 1970s and 80s was their application in the theory of recursive program schemes [6, 7, 21]. Evidently, in this context, the ability of cftg to copy is essential – after all, it is quite a harsh restriction on a program to demand that each formal parameter of a function is used at most once in its body.

In the recent years, there has been renewed interest in cftg in the area of syntax-based natural language processing [14, 15, 16, 18, 20], where tree languages are used to express the linguistic structure of the processed sentences. Here, cftg allow modelling particular linguistic phenomena, which are described as *mildly context-sensitive*. In contrast to recursive program schemes, in this area only non-copying, or *linear*, cftg (l-cftg) are considered, as there is no linguistic motivation for copying, and as the copying power of cftg makes their membership problem computationally hard [22, 25].

The modular design of syntax-based language processing systems requires that the utilized class of tree languages \mathcal{C} possesses a number of closure properties. In particular, for translation tasks it is important that \mathcal{C} is closed under application of linear extended tree transducers (l-xtt). This transducer model was first described by Rounds [23] (under another name), and further investigated, i.a., in [4, 10, 19]. Unfortunately, the closure under l-xtt does *not* hold when \mathcal{C} is the class of context-free tree languages. This is due to a theorem of Arnold and Dauchet, who proved that the context-free tree languages are not closed under inverse linear tree homomorphisms [2]. Trivially, every inverse linear tree homomorphism can be computed by an l-xtt. The proof in [2] works by constructing a *copying* cftg G , and the preimage of the tree language of G under a certain tree homomorphism is shown to be non-context-free.

But since copying is not required anyway – are maybe the *linear* context-free tree languages closed under inverse linear tree homomorphisms? In this work, we answer this question in the negative: there are an l-cftg G_{ex} and a linear tree homomorphism h such that $L = h^{-1}(\mathcal{L}(G_{\text{ex}}))$ is *not* a context-free tree language.¹

The intuition behind our proof is as follows. Every tree t in L is of the form



for some $n \geq 1$ and monadic trees $u_1, v_1, \dots, u_n, v_n$. Here, the root of t is the leftmost symbol σ . The subtrees u_i, v_i , called *chains* in the following, are built up over a parenthesis alphabet, such that the chains u_i contain only opening parentheses, the chains v_i only closing parentheses, and $u_1^R v_1 \dots u_n^R v_n$ is a well-parenthesized word.²

If one were to cut such a tree t into two parts t_1 and t_2 , right through an edge between two σ s, then one could observe that there are some chains u_j in t_1 which contain opening parentheses which are not closed in t_1 , but only in t_2 . A similar observation holds of course for some chains v_j in t_2 . These chains u_j and v_j will be called *critical chains*, and their “unclosed” parts *defects*.

We assume that there is some (not necessarily linear) cftg G with $\mathcal{L}(G) = L$, and show that if G exists, then it can be assumed to be of a special normal form. We analyze the derivations of such a G in normal form. A derivation of a tree t as above begins with a subderivation

$$A(\#, \dots, \#) \Rightarrow_G^* B(s_1, \dots, s_p, \#),$$

where A and B are nonterminals of G , and s_1, \dots, s_p are chains over the parenthesis alphabet. After that, the derivation continues with

$$B(s_1, \dots, s_p, \#) \Rightarrow_G C(s'_1, \dots, s'_p, D(s'_{p+1}, \dots, s'_{2p}, \#)),$$

¹Where, of course, $\mathcal{L}(G_{\text{ex}})$ is the tree language of G_{ex} .

²Here, w^R denotes the reversal of the word w .

for some nonterminals C and D and $s'_1, \dots, s'_{2p} \in \{s_1, \dots, s_p\}$. Finally, C and D derive some terminal trees t_1 and t_2 , respectively. So a derivation of t in G “cuts” t into two pieces as described above!

If G exists, it must therefore prepare the defects of t_1 and t_2 such that they “fit together”, and it can only do so in the initial subderivation $A(\#, \dots, \#) \Rightarrow_G^* B(s_1, \dots, s_p, \#)$. But there are only finitely many arguments of A in which the defects could be prepared. We give a sequence of trees in L such that the number of their defects is strictly increasing, no matter how they are cut apart. Then there is some tree t in this sequence whose defects cannot be prepared fully. Hence it is possible to show by a pumping argument that if $t \in \mathcal{L}(G)$, then there is also a tree $t' \in \mathcal{L}(G)$ whose respective parts do not fit together, and therefore $t' \notin L$. Thus the existence of G is ruled out.

We conclude our work with a positive result: the tree languages of linear *monadic* cftg (lm-cftg), i.e. of l-cftg where each nonterminal has at most one successor, are closed under inverse linear tree homomorphisms. The importance of lm-cftg is underscored by their expressive equivalence to the well-known linguistic formalism of *tree-adjoining grammars* [16, 11]. Our proof is based on the Greibach normal form of lm-cftg [9]. In fact, the closure of Greibach cftg under inverse linear tree homomorphisms was already proven by Arnold and Leguy [5], but their construction results in a copying cftg of higher nonterminal arity.

The article is organized as follows. After establishing some preliminaries in Section 2, we define the tree language L in Section 3. In Section 3.1, the grammar G_{ex} is introduced, while Section 3.2 contains the definition of the homomorphism h and some easy observations on L . In Section 4 we work out a normal form for the assumed cftg G , which allows us to define the concept of *derivation trees* of G in Section 5. This concept facilitates the analysis of the derivations in G . Section 6 contains some properties about factorizations of Dyck words, which formalize the idea of cutting t into two. Finally, in Section 7 we give a counterexample, and rule out the existence of G . Section 8 is about the positive result for lm-cftg.

2 Preliminaries

The set of natural numbers with zero is denoted by \mathbb{N} . For every $m, n \in \mathbb{N}$, the set $\{i \in \mathbb{N} \mid m \leq i \leq n\}$ is denoted by $[m, n]$, and the set $[1, n]$ by $[n]$.

An alphabet is a finite nonempty set. The set of words over A is A^* , the empty word is ε , and $A^+ = A^* \setminus \{\varepsilon\}$. We often abbreviate $\{w\}^*$ by w^* , and $\{w\}^+$ by w^+ , for $w \in A^*$. Let $w = a_1 \cdots a_n$ with $a_1, \dots, a_n \in A$ for some $n \in \mathbb{N}$. Then $|w| = n$, and $w^R = a_n \cdots a_1$, the reversal of w . Moreover, let $v \in A^*$. We say that v is a *factor* of w if there are $w', w'' \in A^*$ such that $w = w'vw''$. If, additionally, $w' = \varepsilon$ (resp. $w'' = \varepsilon$), then v is a *prefix* or *left factor* (resp. *suffix* or *right factor*) of w .

Let A be an alphabet such that $A = B \cup C$, $B \cap C = \emptyset$, and there is a one-to-one relation between the elements of B and C . Define the *Dyck congruence* \equiv to be the smallest congruence relation on A^* such that $bc \equiv \varepsilon$ for each pair of related elements $b \in B$ and $c \in C$. We also say that c *acts as right inverse* to b . The *Dyck language* over A is the set $D_A^* = \{w \in A^* \mid w \equiv \varepsilon\}$. By saying that $w \in A^*$ *reduces to* v (resp. v is the *reduction* of w), we mean that v is the (unique) shortest word in A^* such that $v \equiv w$.

Trees An alphabet Σ equipped with a function $\text{rk}_\Sigma: \Sigma \rightarrow \mathbb{N}$ is a *ranked alphabet*. Let Σ be a ranked alphabet. When Σ is obvious, we write rk instead of rk_Σ . Let $k \in \mathbb{N}$. Then $\Sigma^{(k)} = \text{rk}^{-1}(k)$. We often write $\sigma^{(k)}$ and mean that $\text{rk}(\sigma) = k$.

Let U be a set and Λ denote $\Sigma \cup U \cup C$, where C consists of the three symbols ‘(’, ‘)’, and ‘,’. The set $T_\Sigma(U)$ of *trees (over Σ indexed by U)* is the smallest set $T \subseteq \Lambda^*$ such that $U \subseteq T$, and for every $k \in \mathbb{N}$, $\sigma \in \Sigma^{(k)}$, and $t_1, \dots, t_k \in T$, we also have that $\sigma(t_1, \dots, t_k) \in T$. A tree $\alpha()$, $\alpha \in \Sigma^{(0)}$, is abbreviated by α , a tree $\gamma(t)$, $\gamma \in \Sigma^{(1)}$, by γt , and $T_\Sigma(\emptyset)$ by T_Σ . The notation γt suggests a bijection between Σ^*U and $T_\Sigma(U)$ for monadic ranked alphabets Σ (i.e. $\Sigma = \Sigma^{(1)}$), and in fact we will often confuse such monadic trees with words in writing.

Let $s, t \in T_\Sigma(U)$. The set of *positions (Gorn addresses)* of t is denoted by $\text{pos}(t) \subseteq \mathbb{N}^*$. The number of occurrences of a symbol $\sigma \in \Sigma$ in t is written $|t|_\sigma$. The *size* of t is $|t| = \sum_{\sigma \in \Sigma} |t|_\sigma$. Denote the *label* of t at its position w by $t(w)$, and the *subtree* of t at w by $t|_w$. The result of *replacing* the subtree $t|_w$ in t by s is $t[s]_w$. Fix the infinite set of *variables* $X = \{x_1, x_2, \dots\}$. For each $k \in \mathbb{N}$, let $X_k = \{x_i \mid i \in [k]\}$. Given $n, k \in \mathbb{N}$, $t \in T_\Sigma(X_n)$, and $s_1, \dots, s_n \in T_\Sigma(X_k)$, denote by $t[s_1, \dots, s_n]$ the result of substituting s_i for each occurrence of x_i in t , where $i \in [n]$. Sometimes, especially when no other variable is used, we will write x instead of x_1 .

Magmoids We will heavily use the notation introduced with the concept of *magmoids* [3, 4]. Let $k, n \in \mathbb{N}$. Then the set $\{\langle k, t_1, \dots, t_n \rangle \mid t_1, \dots, t_n \in T_\Sigma(X_k)\}$ is denoted by $T(\Sigma)_k^n$. It is customary to omit the component k from such a tuple, and we will do so from now on.

Let $T(\Sigma) = \bigcup_{n,k \in \mathbb{N}} T(\Sigma)_k^n$. We identify the sets $T_\Sigma(X_k)$ and $T(\Sigma)_k^1$ and write t instead of $\langle t \rangle$. Moreover, we follow the convention of identifying the tree $\sigma(x_1, \dots, x_k) \in T(\Sigma)_k^1$ with the symbol $\sigma \in \Sigma^{(k)}$.

The set of all $u \in T(\Sigma)_k^n$ such that the left-to-right sequence of variables in u is x_1, \dots, x_k , is denoted by $\tilde{T}(\Sigma)_k^n$. The set Θ_k^n of *torsions* is $\{\langle x_{i_1}, \dots, x_{i_n} \rangle \mid i_1, \dots, i_n \in [k]\}$. Note that $\Theta_k^n \subseteq T(\Sigma)_k^n$. We may also understand a torsion $\vartheta \in \Theta_k^n$ as a function $\vartheta: [n] \rightarrow [k]$, such that $\vartheta(i) = j$ if and only if $\vartheta_i = x_j$. The torsion $\langle x_1, \dots, x_n \rangle \in \Theta_n^n$ is denoted by Id_n , and the torsion $\langle x_i \rangle \in T(\Sigma)_k^1$, $i \in [k]$, by π_i^k (when k is clear from the context, we write π_i instead). Then $\pi_i \cdot u$ is the i -th component of the tuple u . A tuple $u \in \tilde{T}(\Sigma)_k^n$ is said to be *torsion-free*. We write $\text{lin}(u)$ for the (unique) tuple $(v, \vartheta) \in \tilde{T}(\Sigma)_m^n \times \Theta_k^n$ such that $u = v \cdot \vartheta$.

Let $n, \ell, k \in \mathbb{N}$, and let $u \in T(\Sigma)_\ell^n$, $v \in T(\Sigma)_k^\ell$. We define $u \cdot v \in T(\Sigma)_k^n$ by

$$u \cdot v = \langle u_1[v_1, \dots, v_\ell], \dots, u_n[v_1, \dots, v_\ell] \rangle.$$

Note that the operation \cdot is associative [12, Prop. 2.4]. If $u \in T(\Sigma)_n^n$, then let $u^0 = \text{Id}_n$ and $u^{(j+1)} = u \cdot u^j$ for every $j \geq 0$. Moreover, given $u \in T(\Sigma)_k^n$ and $v \in T(\Sigma)_k^\ell$, define $[u, v] \in T(\Sigma)_k^{n+\ell}$ by

$$[u, v] = \langle u_1, \dots, u_n, v_1, \dots, v_\ell \rangle.$$

Clearly, this operation is associative, so we will write, e.g., $[u, v, t]$ instead of $[[u, v], t]$.

Trees with a spine We introduce the following special notation which will be helpful to denote the trees we deal with. Let $n, k \in \mathbb{N}$. Then

$$S(\Sigma)_k^n = \{[u, x_{k+1}] \mid u \in T(\Sigma)_k^n\}, \quad \tilde{S}(\Sigma)_k^n = \tilde{T}(\Sigma)_{k+1}^{n+1} \cap S(\Sigma)_k^n, \quad \text{and} \quad \hat{\Theta}_k^n = \Theta_{k+1}^{n+1} \cap S(\Sigma)_k^n.$$

Moreover, for every $s \in T(\Sigma)_{n+1}^1$ and $t \in T(\Sigma)_k^1$, let $s \circ - t = s \cdot [\text{Id}_n, t]$. In an expression containing \cdot and $\circ -$, we assume \cdot to bind stronger than $\circ -$.

Context-free tree grammars A *context-free tree grammar (cftg)* over Σ is a tuple $G = (N, \Sigma, \eta_0, P)$ such that Σ and N are disjoint ranked alphabets (of *terminal* resp. *nonterminal symbols*), $\eta_0 \in T(N \cup \Sigma)_0^1$ (the *axiom*³) and P is a finite set of *productions* of the form $A(x_1, \dots, x_k) \rightarrow t$ for some $k \in \mathbb{N}$, $A \in N^{(k)}$, and $t \in T(N \cup \Sigma)_k^1$. Let $G = (N, \Sigma, S, P)$ be a cftg, and let $n, m \in \mathbb{N}$.

The cftg G is said to be *linear* (resp. *nondeleting*) if in every production $A(x_1, \dots, x_k) \rightarrow t$ in P the right-hand side t contains each variable x_i , $i \in [k]$, at least (resp. at most) once. A linear and nondeleting cftg is *simple*. The cftg G is a *regular tree grammar (rtg)* if $N = N^{(0)}$. Finally, G is *monadic* if $N = N^{(0)} \cup N^{(1)}$. Linear (and monadic) cftg are abbreviated *l-cftg* (*lm-cftg*).

Given $\eta, \zeta \in T(N \cup \Sigma)_m^1$, we write $\eta \Rightarrow_G \zeta$ if there are $A(x_1, \dots, x_k) \rightarrow t$ in P and $\kappa \in T(N \cup \Sigma)_{m+1}^1$, $\tau \in T(N \cup \Sigma)_m^k$ such that κ contains x_{m+1} exactly once,

$$\eta = \kappa \cdot [\text{Id}_m, A \cdot \tau], \quad \text{and} \quad \zeta = \kappa \cdot [\text{Id}_m, t \cdot \tau]. \quad (1)$$

If $\eta, \zeta \in T(N \cup \Sigma)_m^n$ instead, for $n > 1$, write $\eta \Rightarrow_G \zeta$ if there is some $i \in [n]$ such that $\pi_i \cdot \eta \Rightarrow_G \pi_i \cdot \zeta$, and $\pi_j \cdot \eta = \pi_j \cdot \zeta$ for every $j \in [n]$, $j \neq i$. Let $k \in \mathbb{N}$ and $\eta \in T(N \cup \Sigma)_k^1$. Then the set $\{t \in T(\Sigma)_k^1 \mid \eta \Rightarrow_G^* t\}$ is denoted by $\mathcal{L}(G, \eta)$, and the *tree language of G* , denoted by $\mathcal{L}(G)$, is $\mathcal{L}(G, \eta_0)$. We call $L \subseteq T(\Sigma)_0^1$ a *(linear) (monadic) context-free tree language* if there is a (linear) (monadic) cftg G with $L = \mathcal{L}(G)$. Two cftg G and G' are *equivalent* if $\mathcal{L}(G) = \mathcal{L}(G')$.

Recall that there are two restricted modes of derivation for cftg, the OI and the IO mode. In a nutshell, the OI mode requires that a nonterminal may only be rewritten if it occurs outermost in a sentential form. Formally, we write $\eta \xRightarrow{\text{OI}}_G \zeta$ if, additionally to the conditions of (1), the path from the root of κ to the single occurrence of x_{m+1} is only labelled by symbols from $\Sigma \cup X$. The relation $\xRightarrow{\text{OI}}_G$ is extended to $T(N \cup \Sigma)_m^n$ in the same manner as above.

Dually, in the IO mode, a nonterminal may only be rewritten if it occurs innermost. It is well-known that every unrestricted derivation can be emulated by one that is OI, and therefore $\mathcal{L}(G) = \{t \in T(\Sigma)_0^1 \mid \eta_0 \xRightarrow{\text{OI}}_G^* t\}$. Under the IO mode there may indeed be some trees in $\mathcal{L}(G)$ which cannot be derived in this restricted manner [7, 8].

The following lemma fulfills the role of a basic technical lemma on context-free word grammars (e.g. [13, Lemma 3.3.1]). As we must count the number of derivation steps, OI derivations are used.

Lemma 2.1 ([5, Lemma 2]). *Let $G = (N, \Sigma, \eta_0, P)$ be a cftg, $n, p, q, r \in \mathbb{N}$, $\eta \in T(N \cup \Sigma)_q^p$, $\kappa \in T(N \cup \Sigma)_r^q$, and $t \in T(\Sigma)_r^p$. Then $\eta \cdot \kappa \xRightarrow{\text{OI}}_G^n t$ if and only if there are $k, m, \ell \in \mathbb{N}$, $\tilde{u} \in \tilde{T}(\Sigma)_\ell^p$, $\vartheta \in \Theta_q^\ell$, and $v \in T(\Sigma)_r^\ell$ such that*

$$t = \tilde{u} \cdot v, \quad \eta \xRightarrow{\text{OI}}_G^k \tilde{u} \cdot \vartheta, \quad \vartheta \cdot \kappa \xRightarrow{\text{OI}}_G^m v, \quad \text{and} \quad k + m = n.$$

³Just like in the word case, the use of an axiom instead of an initial nonterminal has no impact on the generative power of cftg. But it will be technically convenient to use an axiom.

A cftg G is said to be *total* if $\mathcal{L}(G, A) \neq \emptyset$ for every nonterminal A of G . As the following lemma shows, we may always assume that a cftg is total.

Lemma 2.2 ([1, Appendix]). *For every cftg G , an equivalent total cftg G' can be constructed.*

The proof in [1] assumes that G is in normal form, but with an evident generalization it also goes through without this assumption. The proof's idea is to introduce the production $A \rightarrow \#$, where $\#$ is some dummy symbol, for every non-productive nonterminal A of G , i.e. with $\mathcal{L}(G, A) = \emptyset$. Of course, care must be taken that this dummy symbol is not produced in the course of a derivation in G' which was blocked before in G . Therefore every nonterminal $A \in N^{(k)}$ is annotated with a set $\alpha \subseteq [k]$ of forbidden indices, which prevents choosing a non-productive nonterminal. Apart from this annotation, the construction does not alter the shape of the productions of G .

Tree Homomorphisms Let Σ and Δ be ranked alphabets. A mapping $h: \Sigma \rightarrow \Delta$ is said to be a *tree homomorphism* if $h(\Sigma^{(k)}) \subseteq T_\Delta(X_k)$ for every $k \in \mathbb{N}$. We extend h to a mapping $\hat{h}: T_\Sigma(X) \rightarrow T_\Delta(X)$ by setting $\hat{h}(x_i) = x_i$ for every $i \in \mathbb{N}$ and

$$\hat{h}(\sigma(t_1, \dots, t_k)) = h(\sigma)[\hat{h}(t_1), \dots, \hat{h}(t_k)]$$

for every $k \in \mathbb{N}$, $\sigma \in \Sigma^{(k)}$, and $t_1, \dots, t_k \in T_\Sigma(X)$. In the following, we will no longer distinguish between h and \hat{h} .

We recall the following properties of tree homomorphisms (cf. [5]). Let $h: T_\Sigma(X) \rightarrow T_\Delta(X)$ be a tree homomorphism, and for every $\sigma \in \Sigma^{(k)}$, $k \in \mathbb{N}$, let $h(\sigma) = \tilde{t}_\sigma \cdot \vartheta_\sigma$, where $\tilde{t}_\sigma \in \tilde{T}(\Delta)_\ell^1$ and $\vartheta_\sigma \in \Theta_k^\ell$, for an $\ell \in \mathbb{N}$. We say that h is *linear* (resp. *nondeleting*) if ϑ_σ is injective (resp. surjective), and *alphabetic* or a *delabeling* (*démarquage*) if $\tilde{t}_\sigma \in \Sigma \cup X$, for every $\sigma \in \Sigma$. Moreover, h is *simple* if it is linear and nondeleting. Lastly, h is *elementary ordered* (*élémentaire ordonné*) if there are $\sigma \in \Sigma^{(n)}$, $\delta_1 \in \Delta^{(n-k+1)}$, $\delta_2 \in \Delta^{(k)}$, $n, k \in \mathbb{N}$, and $\ell \in [n+1]$ such that

$$h(\sigma) = \delta_1(x_1, \dots, x_{\ell-1}, \delta_2(x_\ell, \dots, x_{\ell+k-1}), x_{\ell+k}, \dots, x_n)$$

and $h(\omega) = \omega$ for every $\omega \in \Sigma \setminus \{\sigma\}$.

3 The tree language L

We start out by introducing the cftg G_{ex} . The preimage L of $\mathcal{L}(G_{\text{ex}})$ under a simple tree homomorphism h , introduced afterwards, will be shown to be non-context-free later on.

3.1 The grammar G_{ex}

Let $\Delta = \{\delta_1^{(2)}, \delta_2^{(2)}, \#^{(0)}\} \cup \Gamma$, where $\Gamma = \{a^{(1)}, b^{(1)}, c^{(1)}, d^{(1)}\}$. Consider the simple cftg $G_{\text{ex}} = (N_{\text{ex}}, \Delta, \eta_{\text{ex}}, P_{\text{ex}})$ with nonterminal set $N_{\text{ex}} = \{A^{(3)}\}$, axiom $\eta_{\text{ex}} = \delta_1(\#, A(c\#, d\#, \delta_2(\#, \#)))$, and productions in P_{ex} given by

$$\begin{aligned} A(x_1, x_2, x_3) \rightarrow & A(ax_1, bx_2, x_3) \\ & + A(ccx_1, d\#, A(c\#, ddx_2, x_3)) \\ & + \delta_2(cx_1, \delta_1(dx_2, x_3)). \end{aligned}$$

Example 3.1. The following is an example derivation of a tree in $\mathcal{L}(G_{\text{ex}})$.

$$\begin{aligned}
\eta_{\text{ex}} &= \delta_1(\#, x) \circ - A(c\#, d\#, x) \circ - \delta_2(\#, \#) \\
&\Rightarrow_{G_{\text{ex}}}^* \delta_1(\#, x) \circ - A(a^2c\#, b^2d\#, x) \circ - \delta_2(\#, \#) \\
&\Rightarrow_{G_{\text{ex}}} \delta_1(\#, x) \circ - A(c^2a^2c\#, d\#, x) \circ - A(c\#, d^2b^2d\#, x) \circ - \delta_2(\#, \#) \\
&\Rightarrow_{G_{\text{ex}}}^* \delta_1(\#, x) \circ - A(ac^2a^2c\#, bd\#, x) \circ - A(a^2c\#, b^2d^2b^2d\#, x) \circ - \delta_2(\#, \#) \\
&\Rightarrow_{G_{\text{ex}}}^* \delta_1(\#, x) \circ - \delta_2(cac^2a^2c\#, x) \circ - \delta_1(dbd\#, x) \\
&\quad \circ - \delta_2(ca^2c\#, x) \circ - \delta_1(db^2d^2b^2d\#, x) \circ - \delta_2(\#, \#).
\end{aligned}$$

3.2 The homomorphism h and its preimage

Let $\Sigma = \{\sigma^{(3)}, \#^{(0)}\} \cup \Gamma$ and let $h: T_{\Sigma}(X) \rightarrow T_{\Delta}(X)$ be the simple tree homomorphism such that

$$h(\sigma(x_1, x_2, x_3)) = \delta_1(x_1, \delta_2(x_2, x_3)) \quad \text{and} \quad h(\omega) = \omega \text{ for each } \omega \in \Sigma \setminus \{\sigma\}.$$

In the following, we will analyse the tree language $L = h^{-1}(\mathcal{L}(G_{\text{ex}}))$. It is easy to see that every $t \in L$ is of the form

$$\sigma(\#, u_1\#, x) \circ - \sigma(v_1\#, u_2\#, x) \circ - \cdots \circ - \sigma(v_{n-1}\#, u_n\#, x) \circ - \sigma(v_n\#, \#, \#)$$

for some $n \geq 1$, and $u_i \in (ca^*c)^+$, $v_i \in (db^*d)^+$, for each $i \in [n]$. In general, given a tree t of the form

$$\sigma(v_1\#, u_1\#, x) \circ - \cdots \circ - \sigma(v_n\#, u_n\#, \zeta) \quad \text{with } n \geq 1, \quad \zeta \in \{\#\} \cup X, \quad (2)$$

where $v_i \in (db^*d)^+$ and $u_i \in (ca^*c)^+$, $i \in [n]$, we will call the monadic subtrees u_j (resp. v_j) of t the a -chains (resp. the b -chains) of t . A *chain* is either an a - or a b -chain. The rightmost root-to-leaf path in t (that is labeled $\sigma \cdots \sigma \zeta$) will be referred to as t 's *spine*.

For every tree t of the form as in (2), we let $\iota(t) = v_1 u_1^R v_2 u_2^R \cdots v_n u_n^R$. We view Γ as a parenthesis alphabet, such that b acts as right inverse to a , and d to c . Then $\iota(t)$ is a Dyck word, for every $t \in L$.

Proposition 3.1. *For every $t \in L$, $\iota(t) \in D_{\Gamma}^*$.*

Proof. Define ι' for sentential forms of G_{ex} by setting

$$\iota'(\delta_1(v\#, \eta)) = v\iota'(\eta), \quad \iota'(\delta_2(u\#, \eta)) = u^R\iota'(\eta), \quad \iota'(A(v\#, u\#, \eta)) = v u^R \iota'(\eta),$$

and $\iota'(\#) = \varepsilon$, for each $\eta \in T(N_{\text{ex}} \cup \Delta)_0^1$, and $u, v \in \Gamma^*$. One can show by induction for every $\eta \in T(N_{\text{ex}} \cup \Delta)_0^1$ that if $\eta_{\text{ex}} \Rightarrow_{G_{\text{ex}}}^* \eta$, then $\iota'(\eta) \in D_{\Gamma}^*$. Moreover, $\iota'(h(t)) \in D_{\Gamma}^*$ implies that $\iota(t) \in D_{\Gamma}^*$, for any $t \in T(\Delta)_0^1$. This proves the proposition. \square

There is the following relation between the numbers of symbol occurrences in $t \in L$.

Proposition 3.2. *For every $t \in L$, $|t|_c = |t|_d = 4 \cdot |t|_{\sigma} - 6$.*

Proof. One can show by induction for every $\eta \in T(N_{\text{ex}} \cup \Delta)_0^1$ that if $\eta_{\text{ex}} \Rightarrow_{G_{\text{ex}}}^* \eta$, then η fulfills the equation

$$|\eta|_c = |\eta|_d = 2 \cdot |\eta|_{\delta_1} + 2 \cdot |\eta|_{\delta_2} + 3 \cdot |\eta|_A - 6.$$

Obviously, this property transfers to $t \in L$ in the manner described above. \square

Each chain of $t \in L$ is uniquely determined by the other chains of t , because $\iota(t)$ is a Dyck word, and every chain contains either only symbols from $\{a, c\}$, or only symbols from $\{b, d\}$.

Observation 3.1. *Let $t \in L$, let $w \in \text{pos}(t)$ with $t(w) \in \Gamma \cup \{\#\}$, and let $s = t[x_1]_w$. There is exactly one $u \in T(\Gamma \cup \{\#\})_0^1$ such that $s \cdot u \in \mathcal{L}(G_{\text{ex}})$.*

Example 3.2. The preimage of the tree from Example 3.1 under h is

$$t = \sigma(\#, cac^2a^2c\#, x) \circ - \sigma(dbd\#, ca^2c\#, x) \circ - \sigma(db^2d^2b^2d\#, \#, \#).$$

Obviously, $\iota(t) = ca^2c^2acdbdca^2cd b^2d^2b^2d$, and it takes only a little patience to verify that $\iota(t) \in D_\Gamma^*$.

In the following sections, we will prove that there is no cftg G with $\mathcal{L}(G) = L$. Therefore, the following theorem holds.

Theorem 3.1. *The class of linear context-free tree languages is not closed under inverse linear tree homomorphisms.*

4 A normal form for G

Assume there is a cftg $G = (N, \Sigma, \eta_0, P)$ such that $\mathcal{L}(G) = L$. In this section, we show (in a sequence of intermediate normal forms) that if G exists, then it can be chosen to be of a very specific form: Let

$$t = \sigma(u_1\#, v_1\#, x) \circ - \cdots \circ - \sigma(u_n\#, v_n\#, \#) \in L.$$

If we consider the subtrees $\sigma(u_i, v_i)$ as *symbols* from an infinite alphabet Λ , then t can be understood as a word, and L as a word language, over Λ . In fact, in the course of the next lemmas, we will see that therefore G can be assumed to be of a form that is quite close to a context-free word grammar. For example, in Lemma 4.4 it will be shown that the productions of G may be assumed to be of the forms

$$(i) \ A \rightarrow B \cdot u, \text{ with } u \in S(\Gamma)_p^p,$$

$$(ii) \ A \rightarrow B \circ - C, \text{ and}$$

$$(iii) \ A \rightarrow \sigma(x_i, x_j, x_{p+1}),$$

which correspond to (i) chain productions $A \rightarrow B$, (ii) rank 2 productions $A \rightarrow BC$ and (iii) terminal productions $A \rightarrow \sigma$ of context-free grammars. In the next lemma, we start out with distinguishing nonterminals by whether they contribute to the spine of a tree or to its chains.

Lemma 4.1. *We may assume for G that there is $p \in \mathbb{N}$ such that $N = N_s \cup N_c$ for two disjoint sets $N_s = N_s^{(2p)}$ and $N_c = N_c^{(p)}$. Moreover, $\eta_0 = S(\#, \dots, \#)$ for some $S \in N_s$, and every production in P is of one of the following forms:⁴*

$$\begin{aligned} (A1) \quad A &\rightarrow B(C_1, \dots, C_p, D_1, \dots, D_p), & (A4) \quad E &\rightarrow F(C_1, \dots, C_p), \\ (A2) \quad A &\rightarrow x_{p+q}, & (A5) \quad E &\rightarrow x_q, \\ (A3) \quad A &\rightarrow \sigma(x_i, x_j, x_{p+q}), & (A6) \quad E &\rightarrow \gamma(x_q), \end{aligned}$$

where $A, B, D_1, \dots, D_p \in N_s$, $E, F, C_1, \dots, C_p \in N_c$, $i, j, q \in [p]$, and $\gamma \in \Gamma$.

Proof. We begin by assuming that there is a number $p \in \mathbb{N}$ such that $N = N^{(p)}$, the productions in P are of the forms

$$\begin{aligned} (N1) \quad A(x_1, \dots, x_p) &\rightarrow B(C_1(x_1, \dots, x_p), \dots, C_p(x_1, \dots, x_p)), \\ (N2) \quad A(x_1, \dots, x_p) &\rightarrow x_i \text{ for some } i \in [p], \\ (N3) \quad A(x_1, \dots, x_p) &\rightarrow \gamma(x_i) \text{ for some } \gamma \in \Gamma \text{ and } i \in [p], \text{ or} \\ (N4) \quad A(x_1, \dots, x_p) &\rightarrow \sigma(x_i, x_j, x_q) \text{ for some } i, j, q \in [p], \end{aligned}$$

and that $\eta_0 = S(\#, \dots, \#)$ for some $S \in N^{(p)}$. This assumption comes without loss of generality: we may demand that G is in normal form [17, Thm. 14] and then introduce dummy parameters to make every nonterminal of rank p . One fixed parameter x_q can be used to store $\#$ through the course of every derivation, then it is possible to use the production $A \rightarrow x_q$ instead of $A \rightarrow \#$.

Let the regular tree grammar $H = (Q, \Sigma, s, R)$ be given by $Q = \{s, c\}$, and R contains the productions

$$s \rightarrow \sigma(c, c, s) + \# \quad \text{and} \quad c \rightarrow \gamma(c) + \#$$

for every $\gamma \in \Gamma$.

We use Rounds's well-known method [23, 24] to construct a cftg $G' = (N', \Sigma, \eta'_0, P')$ such that $\mathcal{L}(G') = \mathcal{L}(G) \cap \mathcal{L}(H)$. Since $\mathcal{L}(G) \subseteq \mathcal{L}(H)$, it is clear that $\mathcal{L}(G') = \mathcal{L}(G)$. However, as a side-effect of the method, G' is of the desired form. We describe the method briefly, in our own notation. Let $N' = \{A_s^{(2p)} \mid A \in N^{(p)}\} \cup \{A_c^{(p)} \mid A \in N^{(p)}\}$.

Define two functions $\Phi_s: T(N)_p^1 \rightarrow T(N')_{2p}^1$ and $\Phi_c: T(N)_p^1 \rightarrow T(N')_p^1$ by simultaneous induction, such that

$$\Phi_c(x_i) = x_i, \quad \text{and} \quad \Phi_s(x_i) = x_{p+i}$$

for every $x_i \in X_p$, and

$$\begin{aligned} \Phi_c(A(\eta_1, \dots, \eta_p)) &= A_c(\Phi_c(\eta_1), \dots, \Phi_c(\eta_p)), \\ \Phi_s(A(\eta_1, \dots, \eta_p)) &= A_s(\Phi_c(\eta_1), \dots, \Phi_c(\eta_p), \Phi_s(\eta_1), \dots, \Phi_s(\eta_p)) \end{aligned}$$

⁴Recall that $A(x_1, \dots, x_k)$ and $A^{(k)}$ were identified!

for every $A \in N$, and $\eta_1, \dots, \eta_p \in T_N(X_p)$.

For every production $A(x_1, \dots, x_p) \rightarrow \eta$ in P of form (N1) or (N2), the set P' contains the productions $A_c(x_1, \dots, x_p) \rightarrow \Phi_c(\eta)$ and $A_s(x_1, \dots, x_{2p}) \rightarrow \Phi_s(\eta)$. Moreover, for every production in P of form (N3) (resp. (N4)), P' contains the production $A_c(x_1, \dots, x_p) \rightarrow \gamma(\Phi_c(x_i)) = \gamma(x_i)$ (resp. $A_s(x_1, \dots, x_{2p}) \rightarrow \sigma(\Phi_c(x_i), \Phi_c(x_j), \Phi_s(x_q)) = \sigma(x_i, x_j, x_{p+q})$). Let $N_s = \{A_s \mid A \in N\}$ and $N_c = \{A_c \mid A \in N\}$, and let $\eta'_0 = S_s(\#, \dots, \#)$. Then it is easy to see that G' is of the form as demanded above. \square

In the next step we show that we require at most two spine-producing nonterminals on the right-hand side of a production of G . The construction works by guessing beforehand which of the nonterminals of N_s in a production's right-hand side will eventually be chosen to contribute to the tree's spine.

Lemma 4.2. *We may assume for G that there is $p \in \mathbb{N}$ such that $N = N_c \cup N_s$ with $N_c = N^{(p)}$ and $N_s = N^{(p+1)}$. Moreover, $\eta_0 = S(\#, \dots, \#)$ for some $S \in N_s$, and every production of G is of one of the following forms:*

- | | |
|---|--|
| (B1) $A \rightarrow B(C_1, \dots, C_p, x_{p+1}),$ | (B5) $E \rightarrow F(C_1, \dots, C_p),$ |
| (B2) $A \rightarrow B(x_1, \dots, x_p, D),$ | (B6) $E \rightarrow x_i,$ |
| (B3) $A \rightarrow x_{p+1},$ | (B7) $E \rightarrow \gamma(x_i),$ |
| (B4) $A \rightarrow \sigma(x_i, x_j, x_{p+1}),$ | |

where $A, B, D \in N_s$, $E, F, C_1, \dots, C_p \in N_c$, $i, j \in [p]$, and $\gamma \in \Gamma$.

Proof. Assume that $G = (N, \Sigma, \eta_0, P)$ is of the form as given in Lemma 4.1. We will construct an equivalent cftg G'' of the form demanded above.

However, we construct first an intermediate cftg $G' = (N', \Sigma, \eta'_0, P')$, where $N' = N_c \cup N'_s \cup \{S'\}$, such that $S' \notin N_c \cup N'_s$ is a new nonterminal symbol, and

$$N'_s = \{\langle A, q \rangle^{(p+1)} \mid A \in N_s, q \in [p]\}.$$

Moreover, $\eta'_0 = S'(\#, \dots, \#)$, and P' contains the productions

- (i) $\langle A, q \rangle \rightarrow \langle B, \tilde{q} \rangle (C_1, \dots, C_p, \langle D, \tilde{q} \rangle, q)$
for every production of form (A1), and every $q, \tilde{q} \in [p]$;
- (ii) $\langle A, q \rangle \rightarrow x_{p+1}$ for every production of form (A2);
- (iii) $\langle A, q \rangle \rightarrow \sigma(x_i, x_j, x_{p+1})$ for every production of form (A3);
- (iv) every production of form (A4), (A5), or (A6),
- (v) $S' \rightarrow \langle S, q \rangle$ for every $q \in [p]$.

We now prove that $\mathcal{L}(G') = \mathcal{L}(G)$. To this end, it is necessary to consider only OI derivations, as otherwise counting derivation steps becomes bothersome. It is easy to prove by induction for every $n \in \mathbb{N}$, chain-producing nonterminal $C \in N_c$ and $t \in T(\Sigma)_p^1$ that $C \xRightarrow{G}_G^n t$ if and only if $C \xRightarrow{G'}_G^n t$.

Next, we show for every $n \in \mathbb{N}$, $q \in [p]$, $A \in N_s$, and $t \in T(\Sigma)_{p+1}^1$, that

$$A \xRightarrow{G}_G^n t \multimap x_{p+q} \quad \text{if and only if} \quad \langle A, q \rangle \xRightarrow{G'}_G^n t \multimap x_{p+1}.$$

The proof uses the fact that for every $A \in N_s$ and $t \in \mathcal{L}(G, A)$, there is precisely one occurrence of a variable from $\{x_{p+1}, \dots, x_{2p}\}$ in t . We proceed by complete induction on n (using Lemma 2.1 to decompose OI derivations). The base case $n = 0$ holds vacuously. Continue by a case analysis on the production applied first in the derivation. Let $n \in \mathbb{N}$, $A \in N_s$, $q \in [p]$, and $t \in T(\Sigma)_{p+1}^1$. Assume that the production $A \rightarrow B(C_1, \dots, C_p, D_1, \dots, D_p)$ is in P . Then

$$\begin{aligned} & A \xRightarrow{G}_G^n B(C_1, \dots, C_p, D_1, \dots, D_p) \xRightarrow{G}_G^n t \multimap x_{p+q} \\ \text{iff } & \exists m \in \mathbb{N}, \tilde{u} \in \tilde{T}(\Sigma)_m^1, \vartheta \in \Theta_{2p}^m, \nu \in T(\Sigma)_{2p}^m : \\ & B \xRightarrow{G}_G^{n_1} \tilde{u} \cdot \vartheta, \quad \vartheta \cdot [C_1, \dots, C_p, D_1, \dots, D_p] \xRightarrow{G}_G^{n_2} \nu, \quad t \multimap x_{p+q} = \tilde{u} \cdot \nu, \quad n = n_1 + n_2 \\ \text{iff } & \exists m \in \mathbb{N}, \tilde{u} \in \tilde{T}(\Sigma)_{m+1}^1, \vartheta \in \Theta_p^m, \nu \in T(\Sigma)_p^m, \tilde{q} \in [p], w \in T(\Sigma)_{p+1}^1 : \quad (\dagger) \\ & B \xRightarrow{G}_G^{n_1} \tilde{u} \cdot [\vartheta, x_{p+\tilde{q}}], \quad \vartheta \cdot [C_1, \dots, C_p] \xRightarrow{G}_G^{n_2} \nu, \quad D_{\tilde{q}} \xRightarrow{G}_G^{n_3} w \multimap x_{p+q}, \\ & t = u \cdot [\nu, w], \quad n = n_1 + n_2 + n_3 \\ \text{iff } & \exists m \in \mathbb{N}, \tilde{u} \in \tilde{T}(\Sigma)_{m+1}^1, \vartheta \in \Theta_p^m, \nu \in T(\Sigma)_p^m, \tilde{q} \in [p], w \in T(\Sigma)_{p+1}^1 : \\ & \langle B, \tilde{q} \rangle \xRightarrow{G'}_G^{n_1} \tilde{u} \cdot [\vartheta, x_{p+1}], \quad \vartheta \cdot [C_1, \dots, C_p] \xRightarrow{G'}_G^{n_2} \nu, \quad \langle D_{\tilde{q}}, q \rangle \xRightarrow{G'}_G^{n_3} w \multimap x_{p+1}, \\ & t = u \cdot [\nu, w], \quad n = n_1 + n_2 + n_3 \\ \text{iff } & \langle A, q \rangle \xRightarrow{G'}_G^n \langle B, \tilde{q} \rangle (C_1, \dots, C_p, \langle D_{\tilde{q}}, q \rangle) \xRightarrow{G'}_G^n t. \end{aligned}$$

To understand why direction “only if” holds at point (\dagger) above, observe that at this point, $\pi_m \cdot \nu$ has the form $w \multimap x_{p+q}$, for some $w \in T(\Sigma)_{p+1}^1$. Since $\pi_m \cdot \nu$ is generated by $\pi_{\vartheta(m)} \cdot [C_1, \dots, C_p, D_1, \dots, D_p]$, there is some $\tilde{q} \in [p]$ such that $D_{\tilde{q}} \xRightarrow{G}_G^* w \multimap x_{p+q}$.

If the production $A \rightarrow x_{p+q}$ is in P , with $q \in [p]$, then $A \xRightarrow{G}_G^n x_{p+1} \multimap x_{p+q}$ if and only if $\langle A, q \rangle \xRightarrow{G'}_G^n x_{p+1}$ by construction. Finally, if $A \rightarrow \sigma(x_i, x_j, x_{p+q})$ is in P , then $A \xRightarrow{G}_G^n \sigma(x_i, x_j, x_{p+1}) \multimap x_{p+q}$ if and only if $\langle A, q \rangle \xRightarrow{G'}_G^n \sigma(x_i, x_j, x_{p+1})$.

* * *

So for every $t \in T(\Sigma)_{2p}^1$, we have that $t \in \mathcal{L}(G, S)$ if and only if there is some $q \in [p]$ such that $t \multimap x_{p+1} \in \mathcal{L}(G', \langle S, q \rangle)$.

Let $s \in T_\Sigma$. Then $s \in \mathcal{L}(G)$ if and only if there is $t \in \mathcal{L}(G, S)$ such that $s = t \cdot \langle \#, \dots, \# \rangle$, and by the above, this is equivalent to $t \multimap x_{p+1} \in \mathcal{L}(G', \langle S, q \rangle)$ for some $q \in [p]$. By use of the productions (ν) , this holds precisely if $t \multimap x_{p+1} \in \mathcal{L}(G', S')$, i.e. $s \in \mathcal{L}(G')$. Therefore, $\mathcal{L}(G) = \mathcal{L}(G')$.

The cftg G'' results from G' by replacing every production of form $A \rightarrow B(C_1, \dots, C_p, D)$ in P' by the two productions

$$A \rightarrow B_{C_1 \dots C_p}(x_1, \dots, x_p, D) \quad \text{and} \quad B_{C_1 \dots C_p} \rightarrow B(C_1, \dots, C_p, x_{p+1})$$

for some new nonterminal $B_{C_1 \dots C_p}$ of G'' . It is easy to see that $\mathcal{L}(G'') = \mathcal{L}(G')$, so a formal proof is omitted. \square

The next normal form shows that the form of a chain of $t \in L$ is already determined on the spine of t . We can therefore omit chain-producing nonterminals.

Lemma 4.3. *We may assume that G is of the form $G = (N, \Sigma, \eta_0, P)$, such that $N = N^{(p+1)}$ for some $p \in \mathbb{N}$, $\eta_0 = S(\#, \dots, \#)$ for some $S \in N$ and the productions in P are of the forms*

$$\begin{aligned} (C1) \quad & A \rightarrow B \cdot u, \text{ where } u \in S(\Gamma)_p^p, & (C3) \quad & A \rightarrow x_{p+1}, \\ (C2) \quad & A \rightarrow B \circ - C, & (C4) \quad & A \rightarrow \sigma(x_i, x_j, x_{p+1}), \text{ where } i, j \in [p], \end{aligned}$$

and where $A, B, C \in N$.

Proof. Assume that G is of the form given in Lemma 4.2. Moreover, we may assume that $\mathcal{L}(G, E) \neq \emptyset$, by Lemma 2.2. The construction preserves our normal form.

Note that for every $E \in N_c$, we have $\mathcal{L}(G, E) \subseteq T(\Gamma)_p^1$. Choose some fixed tree $u_E \in \mathcal{L}(G, E)$ for each $E \in N_c$, and let $n, m \in \mathbb{N}$. Given $\eta \in T(N \cup \{\#\})_m^n$, we define $\varphi(\eta) \in T(N' \cup \Sigma)_m^n$ as follows. If $n \neq 1$, let $\varphi(\eta) = \langle \varphi(\pi_1 \cdot \eta), \dots, \varphi(\pi_n \cdot \eta) \rangle$. If $n = 1$, let

$$\begin{aligned} \varphi(A \cdot \eta) &= A \cdot \varphi(\eta) & \text{for every } A \in N_s \text{ and } \eta \in T(N)_m^{p+1}, \\ \varphi(E \cdot \eta) &= u_E \cdot \varphi(\eta) & \text{for every } E \in N_c \text{ and } \eta \in T(N)_m^p, \\ \varphi(x_q) &= x_q & \text{for every } q \in [m], \text{ and} \\ \varphi(\#) &= \#. \end{aligned}$$

We construct the cftg $G' = (N', \Sigma, \eta_0, P')$ with $N' = \{A^{(p+1)} \mid A \in N_s\}$, and P' contains the productions

- (i) $A \rightarrow B(\varphi(C_1), \dots, \varphi(C_p), x_{p+1})$ for every production of form (B1) in P ,
- (ii) $A \rightarrow B \circ - D$ for every production of form (B2) in P ,
- (iii) $A \rightarrow x_{p+1}$ for every production of form (B3) in P ,
- (iv) $A \rightarrow \sigma(x_i, x_j, x_{p+1})$ for every production of form (B4) in P .

Observe that in (i), $\varphi(C_i) \in T(\Sigma)_p^1$ for each $i \in [p]$.

(\supseteq) To prove that $\mathcal{L}(G') \subseteq \mathcal{L}(G)$, we show for every $n \in \mathbb{N}$, $A \in N'$, and $t \in T(\Sigma)_{p+1}^1$ that

$$A \Rightarrow_{G'}^n t \quad \text{implies} \quad A \Rightarrow_G^* t.$$

The induction base holds trivially. We continue with the following case analysis. Let $n \in \mathbb{N}$ and $t \in T(\Sigma)_{p+1}^1$.

(I) Let

$$A \Rightarrow_{G'} B(\varphi(C_1), \dots, \varphi(C_p), x_{p+1}) \Rightarrow_{G'}^n t \cdot \langle \varphi(C_1), \dots, \varphi(C_p), x_{p+1} \rangle$$

for some production $A \rightarrow B(C_1, \dots, C_p, x_{p+1})$ in P . By the induction hypothesis, $B \Rightarrow_G^* t$, and clearly $C_i \Rightarrow_G^* \varphi(C_i)$ for each $i \in [p]$, therefore

$$A \Rightarrow_G B \cdot \langle C_1, \dots, C_p, x_{p+1} \rangle \Rightarrow_G^* t \cdot \langle \varphi(C_1), \dots, \varphi(C_p), x_{p+1} \rangle.$$

(II) Let $A \Rightarrow_{G'} B \circ - D \Rightarrow_{G'}^n t$ for some production $A \rightarrow B \circ - D$ in P . Then there are $n_1, n_2 \in \mathbb{N}$, u and $v \in T(\Sigma)_{p+1}^1$ such that

$$B \Rightarrow_{G'}^{n_1} u, \quad D \Rightarrow_{G'}^{n_2} v, \quad n = n_1 + n_2, \quad \text{and} \quad t = u \circ - v.$$

By the induction hypothesis, we have that $B \Rightarrow_G^* u$ and $D \Rightarrow_G^* v$, and therefore

$$A \Rightarrow_G B \circ - D \Rightarrow_G^* u \circ - v.$$

(III) Let $A \Rightarrow_{G'} x_{p+1}$. This means that also $A \Rightarrow_G x_{p+1}$.

(IV) Let $A \Rightarrow_{G'} \sigma(x_i, x_j, x_{p+1})$. Then $A \Rightarrow_G \sigma(x_i, x_j, x_{p+1})$.

* * *

Let $s \in \mathcal{L}(G')$. Then there is some $t \in \mathcal{L}(G', S)$ such that $s = t \cdot \langle \#, \dots, \# \rangle$. By the above, $t \in \mathcal{L}(G, S)$, and therefore $s \in \mathcal{L}(G)$. Thus, $\mathcal{L}(G') \subseteq \mathcal{L}(G)$.

(\subseteq) We continue the proof of correctness with the direction $\mathcal{L}(G) \subseteq \mathcal{L}(G')$. It rests on the following property. For every $n \in \mathbb{N}$, $A \in N_s$, $\eta \in T(N \cup \{\#\})_0^{p+1}$, $s \in \tilde{T}(\Sigma)_1^1$, and $t \in T(\Sigma)_0^1$, if

$$\eta_0 \Rightarrow_G^* s \cdot A \cdot \eta \Rightarrow_G^n s \cdot t, \quad \text{then also} \quad A \cdot \varphi(\eta) \Rightarrow_{G'}^* t.$$

The induction base holds vacuously, so again we continue with a case analysis. Let $n \in \mathbb{N}$, $s \in \tilde{T}(\Sigma)_1^1$, and $t \in T(\Sigma)_0^1$.

(I) Let $\eta_0 \Rightarrow_G^* s \cdot A \cdot \eta \Rightarrow_G s \cdot B(C_1, \dots, C_p, x_{p+1}) \cdot \eta \Rightarrow_G^n s \cdot t$. Then

$$A \cdot \varphi(\eta) \Rightarrow_{G'} B \cdot \langle \varphi(C_1), \dots, \varphi(C_p), x_{p+1} \rangle \cdot \varphi(\eta) = B \cdot \varphi(\langle C_1, \dots, C_p, x_{p+1} \rangle \cdot \eta),$$

and by the induction hypothesis, $B \cdot \varphi(\langle C_1, \dots, C_p, x_{p+1} \rangle \cdot \eta) \Rightarrow_{G'}^* t$.

(II) Let $\eta_0 \Rightarrow_G^* s \cdot A \cdot \eta \Rightarrow_G s \cdot B(x_1, \dots, x_p, D) \cdot \eta \Rightarrow_G^n s \cdot t$. Then

$$A \cdot \varphi(\eta) \Rightarrow_{G'} B(x_1, \dots, x_p, D) \cdot \varphi(\eta) = B \cdot \varphi(\langle x_1, \dots, x_p, D \rangle \cdot \eta),$$

and by the induction hypothesis, $B \cdot \varphi(\langle x_1, \dots, x_p, D \rangle \cdot \eta) \Rightarrow_{G'}^* t$.

(III) Let $\eta_0 \Rightarrow_G^* s \cdot A \cdot \eta \Rightarrow_G s \cdot \pi_{p+1} \cdot \eta \Rightarrow_G^n s \cdot t$ by the production $A \rightarrow x_{p+1}$. Then

$$A \cdot \varphi(\eta) \Rightarrow_{G'} x_{p+1} \cdot \varphi(\eta) = \varphi(\pi_{p+1} \cdot \eta).$$

If $\pi_{p+1} \cdot \eta = \#$, then $\varphi(\pi_{p+1} \cdot \eta) = \# = t$. Otherwise, $\pi_{p+1} \cdot \eta = B \cdot \kappa$ for some $B \in N_s$ and $\kappa \in T(N \cup \{\#\})_0^{p+1}$. By the induction hypothesis, $\varphi(B \cdot \kappa) = B \cdot \varphi(\kappa) \Rightarrow_{G'}^* t$.

(IV) Let $u, v \in \Gamma^*$ such that

$$\eta_0 \Rightarrow_G^* s \cdot A \cdot \eta \Rightarrow_G s \cdot \sigma(\pi_i \cdot \eta, \pi_j \cdot \eta, \pi_{p+1} \cdot \eta) \Rightarrow_G^n s \cdot \sigma(u\#, v\#, t).$$

As in case (III), either $\pi_{p+1} \cdot \eta = \#$, and then $\varphi(\pi_{p+1} \cdot \eta) = t$, or otherwise $\pi_{p+1} \cdot \eta = B \cdot \kappa$ with $B \cdot \varphi(\kappa) \Rightarrow_{G'}^* t$.

Moreover, as $s \cdot \sigma(u\#, v\#, t) \in \mathcal{L}(G)$, Observation 3.1 entails that $\mathcal{L}(G, \pi_i \cdot \eta) = \{u\# \}$ and $\mathcal{L}(G, \pi_j \cdot \eta) = \{v\# \}$, from which we conclude that $\varphi(\pi_i \cdot \eta) = u\#$ and $\varphi(\pi_j \cdot \eta) = v\#$. So

$$A \cdot \varphi(\eta) \Rightarrow_{G'} \sigma(u\#, v\#, \varphi(\pi_{p+1} \cdot \eta)) \Rightarrow_{G'}^* \sigma(u\#, v\#, t).$$

* * *

Let $t \in \mathcal{L}(G)$. Then $\eta_0 = S \cdot \langle \#, \dots, \# \rangle \Rightarrow_G^* t$. The above property entails that

$$S \cdot \varphi(\langle \#, \dots, \# \rangle) = S \cdot \langle \#, \dots, \# \rangle \Rightarrow_{G'}^* t,$$

and hence $t \in \mathcal{L}(G')$. □

It turns out, to derive the spine of $t \in L$, no projecting productions $A \rightarrow x_i$ are required: since G is close to a context-free word grammar with productions (C1) $A \rightarrow B$, (C2) $A \rightarrow BC$, (C3) $A \rightarrow \varepsilon$ and (C4) $A \rightarrow \sigma$, we can eliminate the productions of form (C3) by using the well-known method to remove ε -productions from context-free grammars.

Lemma 4.4. *In Lemma 4.3, it is no restriction to demand that G has no productions of the form (C3).*

Proof. Let $Q = \{A \in N \mid A(x_1, \dots, x_{p+1}) \Rightarrow_G^* x_{p+1}\}$ and construct the cftg $G' = (N, \Sigma, \eta_0, P')$, where P' contains all productions from P of forms (C1), (C2) and (C4). Moreover, for every production of form (C2), P' contains the productions

$$A \rightarrow B \quad \text{if} \quad C \in Q, \quad \text{and} \quad A \rightarrow C \quad \text{if} \quad B \in Q.$$

Observe that both productions are of form (C1).

We prove that $\mathcal{L}(G') = \mathcal{L}(G)$. For the direction $\mathcal{L}(G') \subseteq \mathcal{L}(G)$, we show for every $n \in \mathbb{N}$, $A \in N$, and $t \in T(\Sigma)_{p+1}^1$, that if $A \Rightarrow_{G'}^n t$, then also $A \Rightarrow_G^* t$. The proof is by complete induction on n . The induction base is trivial; we proceed by a case analysis on the form of the production applied first.

Assume that $A \Rightarrow_{G'} B \Rightarrow_{G'}^n t$. From the induction hypothesis, $B \Rightarrow_G^* t$. There are three subcases. Either, the production $A \rightarrow B$ is in P , in which case $A \Rightarrow_G^* t$. Otherwise, by

construction, there is some production $A \rightarrow B \multimap C$ or $A \rightarrow C \multimap B$ in P such that $C \Rightarrow_G^* x_{p+1}$. If it is $A \rightarrow B \multimap C$ (the other case is analogous), we have

$$A \Rightarrow_G B \multimap C \Rightarrow_G^* B \Rightarrow_G^* t.$$

The property is trivially true if the applied production is from P . For every $s \in \mathcal{L}(G')$, there is some $t \in \mathcal{L}(G', S)$ with $s = t \cdot \langle \#, \dots, \# \rangle$. By the above, $S \Rightarrow_G^* t$, and therefore $s \in \mathcal{L}(G)$.

* * *

It remains to show the direction $\mathcal{L}(G) \subseteq \mathcal{L}(G')$. We show for every $n \in \mathbb{N}$, $A \in N$, and $t \in T_\Sigma(X_{p+1})$, that if $A \Rightarrow_G^n t$ and $t \neq x_{p+1}$, then also $A \Rightarrow_{G'}^* t$. The induction base is trivial, so again we continue by a case analysis on derivations of nonzero length. Let $n \in \mathbb{N}$, $A \in N$, and $t \in T_\Sigma(X_{p+1})$ with $t \neq x_{p+1}$.

If $A \Rightarrow_G B \multimap C \Rightarrow_G^n t$, then there are $n_1, n_2 \in \mathbb{N}$ and $t_1, t_2 \in T(\Sigma)_{p+1}^1$ with $t = t_1 \multimap t_2$, $B \Rightarrow_G^{n_1} t_1$, $C \Rightarrow_G^{n_2} t_2$, and $n = n_1 + n_2$. If neither $t_1 = x_{p+1}$ nor $t_2 = x_{p+1}$, then by the induction hypothesis, also

$$A \Rightarrow_{G'} B \multimap C \Rightarrow_{G'}^* t_1 \multimap t_2 = t.$$

If precisely one of t_1 and t_2 is equal to x_{p+1} (say $t_1 = x_{p+1}$, the other case is analogous), then the production $A \rightarrow C$ is in P' . So, with the induction hypothesis,

$$A \Rightarrow_{G'} C \Rightarrow_{G'}^* t_2 = t.$$

The case $t_1 = t_2 = x_{p+1}$ is precluded by the assumption that $t \neq x_{p+1}$.

Similarly, the case that the first production is of form (C3) is precluded by the assumption on t . For any other production, the proof goes through without surprises.

Now let $s \in \mathcal{L}(G)$. Then there is $t \in \mathcal{L}(G, S)$ such that $s = t \cdot \langle \#, \dots, \# \rangle$. Note that $t \neq x_{p+1}$, because $\# \notin \mathcal{L}(G)$. Thus by the above property, $S \Rightarrow_{G'}^* t$, and therefore $s \in \mathcal{L}(G')$. \square

Finally, it is convenient to remove the torsions from productions of the form (C1). Then whenever $A \Rightarrow_G^* B \cdot u$, we know that u is torsion-free. The construction works by guessing which torsion will be applied in the next derivation step, and pre-arranging this torsion in the tuple of the current production. However, there is a price to pay: we must now allow for torsions in “branching” productions $A \rightarrow B \cdot \vartheta_1 \multimap C \cdot \vartheta_2$.

Lemma 4.5. *We may assume that G is of the form $G = (N, \Sigma, \eta_0, P)$, such that $N = N^{(p+1)}$ for some $p \in \mathbb{N}$, $\eta_0 = S(\#, \dots, \#)$ for some $S \in N$ and the productions in P are of the forms*

$$(D1) \ A \rightarrow B \cdot u, \text{ where } u \in \widetilde{S}(\Gamma)_p^p,$$

$$(D2) \ A \rightarrow B \cdot \vartheta_1 \multimap C \cdot \vartheta_2, \text{ where } \vartheta_1, \vartheta_2 \in \widehat{\Theta}_p^p,$$

$$(D3) \ A \rightarrow \sigma(x_i, x_j, x_{p+1}), \text{ where } i, j \in [p],$$

and where $A, B, C \in N$.

Proof. Assume that G is as in Lemma 4.4. Construct a new cftg $G' = (N', \Sigma, \eta'_0, P')$, where $N' = N \times \widehat{\Theta}_p^p \cup \{S'\}$ for some distinct nonterminal S' , $\eta'_0 = S'(\#, \dots, \#)$, and P' contains the productions

- (i) $A^{\vartheta'} \rightarrow B^{\vartheta} \cdot s$ for every production of form (C1) and $\vartheta \in \widehat{\Theta}_p^p$, where $\text{lin}(\vartheta \cdot u) = (s, \vartheta')$;
- (ii) $A^{\text{Id}_{p+1}} \rightarrow B^{\vartheta_1} \cdot \vartheta_1 \multimap C^{\vartheta_2} \cdot \vartheta_2$ for every production of form (C2), and $\vartheta_1, \vartheta_2 \in \widehat{\Theta}_p^p$;
- (iii) $A^{\text{Id}_{p+1}} \rightarrow \sigma(x_i, x_j, x_{p+1})$ for every production of form (C4);
- (iv) $S' \rightarrow S^{\vartheta}$ for every $\vartheta \in \widehat{\Theta}_p^p$.

To prove the construction correct, we demonstrate for every $n \in \mathbb{N}$, $A \in N$, $v \in S(\Gamma)_p^p$, and $t \in T(\Sigma)_{p+1}^1$, that

$$A \cdot v \Rightarrow_G^n t \quad \text{if and only if} \quad \exists \vartheta \in \widehat{\Theta}_p^p : A^{\vartheta} \cdot \vartheta \cdot v \Rightarrow_{G'}^n t.$$

The proof is by complete induction on n . The induction base holds trivially, hence we proceed by a case analysis on derivations of nonzero length. Assume therefore that $n \in \mathbb{N}$, $A \in N$, $v \in S(\Gamma)_p^p$, and $t \in T(\Sigma)_{p+1}^1$.

(I) By construction, $A \cdot v \Rightarrow_G \sigma(\pi_i \cdot v, \pi_j \cdot v, x_{p+1})$ if and only if $A^{\text{Id}_{p+1}} \cdot v \Rightarrow_{G'} \sigma(\pi_i \cdot v, \pi_j \cdot v, x_{p+1})$.

(II) Assume that $A \cdot v \Rightarrow_G B \cdot v \multimap C \cdot v \Rightarrow_G^n t$. Then there are $n_1, n_2 \in \mathbb{N}$, t_1 , and $t_2 \in T(\Sigma)_{p+1}^{p+1}$ such that

$$B \cdot v \Rightarrow_G^{n_1} t_1, \quad C \cdot v \Rightarrow_G^{n_2} t_2, \quad t = t_1 \multimap t_2, \quad \text{and} \quad n = n_1 + n_2.$$

By induction, there are $\vartheta_1, \vartheta_2 \in \widehat{\Theta}_p^p$ such that $B^{\vartheta_1} \cdot \vartheta_1 \cdot v \Rightarrow_{G'}^{n_1} t_1$ and $C^{\vartheta_2} \cdot \vartheta_2 \cdot v \Rightarrow_{G'}^{n_2} t_2$. Thus,

$$A^{\text{Id}_{p+1}} \cdot v \Rightarrow_{G'} B^{\vartheta_1} \cdot \vartheta_1 \cdot v \multimap C^{\vartheta_2} \cdot \vartheta_2 \cdot v \Rightarrow_{G'}^{n_1+n_2} t_1 \multimap t_2 = t.$$

Conversely, let $\vartheta_1, \vartheta_2, \vartheta_3 \in \widehat{\Theta}_p^p$, and $n_1, n_2 \in \mathbb{N}$ such that

$$A^{\vartheta_1} \cdot \vartheta_1 \cdot v \Rightarrow_{G'} B^{\vartheta_2} \cdot \vartheta_2 \cdot \vartheta_1 \cdot v \multimap C^{\vartheta_3} \cdot \vartheta_3 \cdot \vartheta_1 \cdot v \Rightarrow_{G'}^n t.$$

By construction, $\vartheta_1 = \text{Id}_{p+1}$. Moreover, there are $n_1, n_2 \in \mathbb{N}$, t_1 , and $t_2 \in T(\Sigma)_{p+1}^{p+1}$ such that

$$B^{\vartheta_2} \cdot \vartheta_2 \cdot v \Rightarrow_{G'}^{n_1} t_1, \quad C^{\vartheta_3} \cdot \vartheta_3 \cdot v \Rightarrow_{G'}^{n_2} t_2, \quad t = t_1 \multimap t_2, \quad \text{and} \quad n = n_1 + n_2.$$

By the induction hypothesis, $B \cdot v \Rightarrow_G^{n_1} t_1$ and $C \cdot v \Rightarrow_G^{n_2} t_2$, thus

$$A \cdot v \Rightarrow_G B \cdot v \multimap C \cdot v \Rightarrow_G^{n_1+n_2} t_1 \multimap t_2 = t.$$

(III) Assume finally that $A \cdot v \Rightarrow_G B \cdot u \cdot v \Rightarrow_G^n t$ for some $n \in \mathbb{N}$. By the induction hypothesis, there is some $\vartheta \in \widehat{\Theta}_p^p$ such that $B^\vartheta \cdot \vartheta \cdot u \cdot v \Rightarrow_{G'}^n t$. By construction, there is a production $A^{\vartheta'} \rightarrow B^\vartheta \cdot s$, where $s \in \widetilde{S}(\Gamma)_p^p$ with $s \cdot \vartheta' = \vartheta \cdot u$. Thus we have

$$A^{\vartheta'} \cdot \vartheta' \cdot v \Rightarrow_{G'} B^\vartheta \cdot s \cdot \vartheta' \cdot v = B^\vartheta \cdot \vartheta \cdot u \cdot v \Rightarrow_{G'}^n t.$$

For the other direction, let $A^{\vartheta'} \cdot \vartheta' \cdot v \Rightarrow_{G'} B^\vartheta \cdot s \cdot \vartheta' \cdot v \Rightarrow_{G'}^n t$ for some $n \in \mathbb{N}$, ϑ and $\vartheta' \in \widehat{\Theta}_p^p$. By construction, there is some production $A \rightarrow B \cdot u$ in P , such that $s \cdot \vartheta' = \vartheta \cdot u$. Hence, $B^\vartheta \cdot s \cdot \vartheta' \cdot v = B^\vartheta \cdot \vartheta \cdot u \cdot v \Rightarrow_{G'}^n t$. By the induction hypothesis, $B \cdot u \cdot v \Rightarrow_G^n t$. Thus also $A \cdot v \Rightarrow_G B \cdot u \cdot v \Rightarrow_G^n t$.

* * *

Let $t \in T(\Sigma)_0^1$. Then $t \in \mathcal{L}(G)$ if and only if $t \in \mathcal{L}(G, S \cdot \langle \#, \dots, \# \rangle)$. By the above property, this holds precisely if there is some $\vartheta \in \widehat{\Theta}_k^k$ such that $t \in \mathcal{L}(G, S^\vartheta \cdot \vartheta \cdot \langle \#, \dots, \# \rangle)$, and it is easy to see that $S^\vartheta \cdot \vartheta \cdot \langle \#, \dots, \# \rangle = S^\vartheta \cdot \langle \#, \dots, \# \rangle$. By construction of G' , we receive that

$$t \in \mathcal{L}(G) \quad \text{iff} \quad t \in \mathcal{L}(G, S \cdot \langle \#, \dots, \# \rangle) \quad \text{iff} \quad t \in \mathcal{L}(G', S' \cdot \langle \#, \dots, \# \rangle) \quad \text{iff} \quad t \in \mathcal{L}(G'),$$

and therefore $\mathcal{L}(G) = \mathcal{L}(G')$. □

Assume for the rest of this work that there is a cftg G of the form stated in Lemma 4.5 such that $\mathcal{L}(G) = L$. Let χ denote the tuple $\langle \#, \dots, \# \rangle$. Then $\eta_0 = S \cdot \chi$.

5 Derivation trees

A derivation of a tree $t \in \mathcal{L}(G)$ can be described faithfully by a binary tree κ .⁵ These *derivation trees* will help us analyze the structure of the derivations in G .

Formally, let κ be a binary tree such that each position $\delta \in \text{pos}(\kappa)$ is equipped with two nonterminal symbols A_δ and $B_\delta \in N$, a torsion-free tuple $s_\delta \in \widetilde{S}(\Gamma)_p^p$, a torsion $\vartheta_\delta \in \widehat{\Theta}_p^p$, and two numbers i_δ and $j_\delta \in [p]$. Then κ is an $(A_\varepsilon, \vartheta_\varepsilon)$ -*derivation tree* if for every $\delta \in \text{pos}(\kappa)$,

- (i) $A_\delta \Rightarrow_G^* B_\delta \cdot s_\delta$,
- (ii) if δ is a leaf of κ , then the production $B_\delta \rightarrow \sigma(x_{i_\delta}, x_{j_\delta}, x_{p+1})$ is in P ,
- (iii) otherwise, $B_\delta \rightarrow A_{\delta_1} \cdot \vartheta_{\delta_1} \multimap A_{\delta_2} \cdot \vartheta_{\delta_2}$ is a production in P .

Let $t \in T(\Sigma)_{p+1}^1$. We say that κ is an $(A_\varepsilon, \vartheta_\varepsilon)$ -*derivation tree of t* (or: κ *derives t*) if either κ has only one node and $t = \sigma(x_{i_\varepsilon}, x_{j_\varepsilon}, x_{p+1}) \cdot s_\varepsilon \cdot \vartheta_\varepsilon$, or, otherwise, there are $t_1, t_2 \in T(\Sigma)_{p+1}^1$ such that $\kappa|_1$ derives t_1 , $\kappa|_2$ derives t_2 , and $t = (t_1 \multimap t_2) \cdot s_\varepsilon \cdot \vartheta_\varepsilon$. An (S, Id_{p+1}) -*derivation tree (of t)* will simply be called a *derivation tree (of t)*.

There is the following relation between derivations and derivation trees.

Proposition 5.1. *Let $t \in T(\Sigma)_{p+1}^1$, let $A \in N$, and $\vartheta \in \widehat{\Theta}_p^p$. Then $A \cdot \vartheta \Rightarrow_G^* t$ if and only if there is an (A, ϑ) -derivation tree of t .*

⁵As a subset $\text{pos}(\kappa) \subseteq \{1, 2\}^*$ that is prefix-closed and such that $w1 \in \text{pos}(\kappa)$ iff $w2 \in \text{pos}(\kappa)$.

Proof. Straightforward by complete induction on $|t|_\sigma$. \square

As a direct corollary, $t \cdot \chi \in L$ if and only if there is a derivation tree of t . We close our discussion of derivation trees with the following pumping lemma. It states that if there is some s_δ in κ which has a sufficiently large component, then an iterable pair of nonterminals occurs in the derivation of s_δ .

In the sequel, fix the pumping number $H = |N| \cdot h_{\max}$, where h_{\max} is the maximal size of a component of u in a production of G of form (D1).

Lemma 5.1. *Let κ be a derivation tree and $\delta \in \text{pos}(\kappa)$. If there are $i \in [p]$ and $w, w' \in \Gamma^*$ such that $\pi_i \cdot s_\delta = w'wx_i$ and $|w| > H$, then there exist $v, y, z \in \tilde{S}(\Sigma)_p^p$ such that*

- (i) $s_\delta = v \cdot y \cdot z$,
- (ii) $\pi_i \cdot y \cdot z$ is a suffix of wx_i ,
- (iii) $|\pi_i \cdot y| > 0$, and
- (iv) for each $j \in \mathbb{N}$, $A_\delta \Rightarrow_G^* B_\delta \cdot v \cdot y^j \cdot z$.

Proof. By definition of κ , $A_\delta \Rightarrow_G^* B_\delta \cdot s_\delta$. So there are $n \in \mathbb{N}$, $C_1, \dots, C_n \in N$ and $e^{(1)}, \dots, e^{(n)} \in \tilde{S}(\Gamma)_p^p$ such that

$$C_1 \cdot e^{(1)} \Rightarrow_G C_2 \cdot e^{(2)} \cdot e^{(1)} \Rightarrow_G \dots \Rightarrow_G C_n \cdot e^{(n)} \dots e^{(1)}$$

where $C_1 = A_\delta$, $C_n = B_\delta$, $e^{(1)} = \text{Id}_{p+1}$, and $e^{(n)} \dots e^{(1)} = s_\delta$.

If C_1, \dots, C_n are pairwise distinct, then $n \leq |N|$ and the maximal size of a component of s_δ is $|N| \cdot h_{\max} = H$, which contradicts the assumption that $|w| > H$. We can therefore choose two indices $\ell, k \in [q]$ with $\ell < k$ such that $C_\ell = C_k$, the size of $\pi_i \cdot e^{(k)} \dots e^{(\ell+1)}$ is nonzero, and ℓ and k are the two smallest numbers with these properties. Let

$$v = e^{(n)} \dots e^{(k+1)}, \quad y = e^{(k)} \dots e^{(\ell+1)}, \quad \text{and} \quad z = e^{(\ell)} \dots e^{(1)}.$$

Then for every $j \in \mathbb{N}$,

$$A_\delta \cdot \text{Id}_{p+1} \Rightarrow_G^* C_\ell \cdot z \Rightarrow_G^* C_k \cdot y^j \cdot z \Rightarrow_G^* B_\delta \cdot v \cdot y^j \cdot z.$$

Moreover, the size of $\pi_i \cdot y \cdot z$ is at most H , therefore $\pi_i \cdot y \cdot z$ is a suffix of wx_i . \square

6 Dyck words and sequences of chains

This section prepares some necessary notions for the upcoming counterexample. We introduce a sequence U_1, U_2, \dots of Dyck words. Later, an element of this sequence will contribute to the chains of the tree t used in the counterexample. As described in the introduction, the proof revolves around the factorization of t into trees t_1 and t_2 that is induced by the derivation of t . So we will analyze the corresponding factorizations of the Dyck words U_i .

Moreover, we will introduce here the notion of *defects*, which can be understood as the “unclosed parentheses” in t_1 , resp. t_2 . Finally, a lemma on *perturbations* is given, which will be used to show that if the defects in t_1 are modified (or: perturbed), then the word formed

by the chains of the resulting tree lies in another Dyck congruence class. This implies that the resulting tree does not “fit together” with t_2 any longer.

First of all, let us fix the following constants. Let $q = 2p$, and let $m = 2^{q-1} + 1$. For every $i \in \mathbb{N}$, let $\alpha_i = ca^{imH}c$ and $\beta_i = db^{imH}d$. Define the sequence U_1, U_2, \dots of words over Γ by

$$U_1 = \alpha_1\beta_1 \quad \text{and} \quad U_{i+1} = \alpha_{i+1}U_iU_i\beta_{i+1} \quad \text{for every } i \geq 1.$$

We make the following observation.

Observation 6.1. *For every $i \geq 1$,*

(i) $U_i \in D_\Gamma^*$, and

(ii) $U_i = u_1v_1 \cdots u_nv_n$, where $n = 2^{i-1}$, $u_j \in (ca^+c)^+$, and $v_j \in (db^+d)^+$, for $j \in [n]$.

For each U_i of the above form, let $Z_i = \langle u_1^R, v_1, \dots, u_n^R, v_n \rangle$. The components u_ℓ^R and v_ℓ of Z_i will also be called *chains*, as later on they will end up as the chains of some $t \in L$. For every factorization of Z_i into

$$Z'_i = \langle u_1^R, v_1, u_2^R, v_2, \dots, u_j^R \rangle \quad \text{and} \quad Z''_i = \langle v_j, u_{j+1}^R, v_{j+1}, \dots, u_n^R, v_n \rangle, \quad j \in [n],$$

consider the respective factors $P_{i,j} = u_1v_1u_2v_2 \cdots u_j$ and $S_{i,j} = v_ju_{j+1}v_{j+1} \cdots u_nv_n$ of U_i .

Proposition 6.1. *The factors $P_{i,j}$ and $S_{i,j}$ can be written as*

$$P_{i,j} = \alpha_i V_{i-1} \alpha_{i-1} \cdots V_1 \alpha_1 \quad \text{and} \quad S_{i,j} = \beta_1 W_1 \cdots \beta_{i-1} W_{i-1} \beta_i, \quad (3)$$

such that $V_\ell, W_\ell \in \{\varepsilon, U_\ell\}$ and $V_\ell \neq W_\ell$ for every $\ell \in [i-1]$.

Proof. By induction on i . The base case $U_1 = \alpha_1\beta_1$ has only one factorization, $P_{1,1} = \alpha_1$ and $S_{1,1} = \beta_1$, which fulfills the property. Let $i \geq 1$ and consider $U_{i+1} = \alpha_{i+1}U_iU_i\beta_{i+1}$. A factorization $P_{i+1,j}S_{i+1,j}$ of U_{i+1} induces a factorization of either the first or the second occurrence of U_i into, say $P_{i,j'}$ and $S_{i,j'}$ for some $j' \in [2^{i-1}]$. Therefore, $U_{i+1} = \alpha_{i+1}V_iP_{i,j'}S_{i,j'}W_i\beta_{i+1}$ for $V_i, W_i \in \{\varepsilon, U_i\}$ with $V_i \neq W_i$. By induction, $P_{i,j'} = \alpha_i V_{i-1} \alpha_{i-1} \cdots V_1 \alpha_1$, and therefore $P_{i+1,j} = \alpha_{i+1}V_i \alpha_i V_{i-1} \alpha_{i-1} \cdots V_1 \alpha_1$, for V_i, \dots, V_1 as given above. The same kind of argument works for $S_{i+1,j}$. \square

Assume a factorization of U_i into $P_{i,j}$ and $S_{i,j}$ as given in (3). Then we denote by $D_{i,j}$ the word

$$\$ \alpha_i V'_{i-1} \alpha_{i-1} \cdots V'_1 \alpha_1 \$ \beta_1 W'_1 \cdots \beta_{i-1} W'_{i-1} \beta_i \$$$

over $\Gamma \cup \{\$, \}$, where for every $\ell \in [i-1]$, $V'_\ell = \$$ if $V_\ell = U_\ell$, and $V'_\ell = \varepsilon$ if $V_\ell = \varepsilon$, and, analogously, $W'_\ell = \$$ if $W_\ell = U_\ell$, and $W'_\ell = \varepsilon$ if $W_\ell = \varepsilon$. Let $\ell, k \in \mathbb{N}$ with $\ell \leq k$. We say that a word $\gamma = \alpha_\ell \cdots \alpha_k$ (resp. $\gamma = \beta_\ell \cdots \beta_k$) is an *a-defect* (resp. a *b-defect*) in $D_{i,j}$ if $\$ \gamma^R \$$ (resp. $\$ \gamma \$$) occurs in $D_{i,j}$. When the factorization is clear, the reference to $D_{i,j}$ is omitted. Both *a-defects* and *b-defects* will be called *defects*. A chain in Z_i whose suffix is a defect is called a *critical chain*.

Proposition 6.2. *Consider a factorization of U_i into $P_{i,j}$ and $S_{i,j}$.*

- (1) There is no $\ell \in [i]$ such that α_ℓ (or β_ℓ) occurs in two distinct defects.
- (2) The number of defects in $D_{i,j}$ is $i + 1$.
- (3) Each a -defect (resp. b -defect) is the suffix of some chain u_n (resp. v_n) in Z_i , with $n \in [2^{i-1}]$.

Proof. For (1), observe that the a -defects in $D_{i,j}$ are disjoint (non-overlapping) factors of the word $\alpha_1 \cdots \alpha_i$. A similar observation can be made for the b -defects in $D_{i,j}$. For (2), it is easy to see from Proposition 6.1 that there are exactly $i + 2$ occurrences of the symbol $\$$ in $D_{i,j}$. So there are $i + 1$ factors of the form $\$ \gamma \$$ in $D_{i,j}$, for $\gamma \in \Gamma^*$. By (1), the defects are pairwise distinct, so $D_{i,j}$ contains precisely $i + 1$ defects.

Regarding (3), let $\gamma = \alpha_\ell \cdots \alpha_k$, $k \geq \ell$, be an a -defect in $D_{i,j}$ and let

$$D_{i,j} = D' \$ \underbrace{\alpha_k \cdots \alpha_\ell}_{\gamma^R} \$ D'' \quad \text{for some } D', D'' \in (\Gamma \cup \{\$\})^*.$$

By definition of $D_{i,j}$, $P_{i,j}$ is of the form

$$P_{i,j} = P' U_k \alpha_k \cdots U_\ell \alpha_\ell P'' \quad \text{for some } P', P'' \in \Gamma^*$$

if $k < i$, and $P_{i,j} = \alpha_k \cdots U_\ell \alpha_\ell P''$ if $k = i$. As U_k ends with β_k , γ is the suffix of some chain u_n in Z_i . A similar argument can be made if γ is a b -defect. \square

Let $P, P' \in (ca^*c)^*$. We say that P' is a *perturbation* of P if it results from P by modifying the exponents of a in P . More precisely, let P be of the form

$$P = w_0 a^{f_1} w_1 \cdots w_{\ell-1} a^{f_\ell} w_\ell,$$

such that $\ell \in \mathbb{N}$, $w_0, \dots, w_\ell \in c^*$, and for each $i \in [\ell]$, $f_i > 0$. Then $P' \in \Gamma^*$ is called a *perturbation* of P if

$$P' = w_0 a^{f'_1} w_1 \cdots w_{\ell-1} a^{f'_\ell} w_\ell,$$

for some $f'_1, \dots, f'_\ell \in \mathbb{N}$. The only perturbation of ε is ε itself.

Lemma 6.1. Consider a factorization of U_i into $P_{i,j}$ and $S_{i,j}$, and let $P'_{i,j}$ be a perturbation of $P_{i,j}$, i.e.

$$P_{i,j} = \alpha_i V_{i-1} \alpha_{i-1} \cdots V_1 \alpha_1 \quad \text{and} \quad P'_{i,j} = \alpha'_i V'_{i-1} \alpha'_{i-1} \cdots V'_1 \alpha'_1. \quad (4)$$

Then $P'_{i,j} \equiv P_{i,j}$ if and only if $V'_\ell \equiv \varepsilon$ for every $\ell \in [i-1]$ and $\alpha'_\ell = \alpha_\ell$ for every $\ell \in [i]$.

Proof. The direction “if” is trivial. For the other direction, we first prove for every $i > 0$ and every perturbation U'_i of U_i that either $U'_i \equiv \varepsilon$ or $U'_i = cXd$ for some $X \not\equiv \varepsilon$. The proof is by induction on i . For the base case, consider a perturbation $U'_1 = ca^p cdb^q d$ of U_1 , where $p, q \in \mathbb{N}$. Since $U'_1 \equiv ca^p b^q d$, $U'_1 \not\equiv \varepsilon$ implies that $p \neq q$, and therefore $a^p b^q \not\equiv \varepsilon$. Assume now a perturbation

$$P'_{i+1} = ca^p c P'_i P''_i d b^q d, \quad p, q \in \mathbb{N},$$

of P_{i+1} , where P'_i and P''_i are perturbations of P_i . If $P_{i+1} \not\equiv \varepsilon$, then either $P'_i P''_i \equiv \varepsilon$ and $p \neq q$ as above. Otherwise $P'_i P''_i$ is of the form cXd with $X \not\equiv \varepsilon$. But then P'_{i+1} is also of this form.

Algorithm 1 Derivation of $h(t)$ in G_{ex}

```
 $\eta \leftarrow \eta_{\text{ex}}$ 
for all  $j \in \{1, \dots, q\}$  do
  repeat  $j \cdot m \cdot H$  times
     $\eta \leftarrow \text{apply}(A \rightarrow A(ax_1, bx_2, x_3), \eta)$ 
  end repeat
   $\eta \leftarrow \text{apply}(A \rightarrow A(ccx_1, d\#, A(c\#, ddx_2, x_3)), \eta)$ 
end for
 $\eta \leftarrow \text{apply}(A \rightarrow \delta_2(cx_1, \delta_1(dx_2, x_3)), \eta)$ 
```

We can now prove the direction “only if”. Let $P'_{i,j} \equiv P_{i,j}$. As $V_\ell \in \{U_\ell, \varepsilon\}$ for every $\ell \in [i-1]$, $P_{i,j}$ reduces to $\alpha_i \cdots \alpha_1$. Assume that there is some $\ell \in [i-1]$ with $V'_\ell \neq \varepsilon$. Then the reduction of $P'_{i,j}$ would contain an occurrence of c , by the property shown above. But this is in contradiction to the assumption that $P'_{i,j} \equiv P_{i,j}$. Hence, $V'_1, \dots, V'_{i-1} \equiv \varepsilon$. Then clearly also $\alpha'_\ell = \alpha_\ell$ for every $\ell \in [i]$. \square

Let us remark that an analogous lemma can be formulated for perturbations of $S_{i,j}$. However, we will only consider perturbations of $P_{i,j}$ afterwards.

7 A witness for $\mathcal{L}(G) \neq L$

In this section, we choose a tree $t \in L$ whose chains form a sufficiently large word U_i . By viewing a derivation tree κ of t , which induces a factorization $t = t_1 \circ t_2$, we will see that the pumping lemma from Section 5 can be applied, and this leads to a perturbation in the defects of t_1 . By Lemma 6.1 right above, we receive the desired contradiction.

Let $Z_q = \langle u_1, v_1, \dots, u_{m-1}, v_{m-1} \rangle$, recalling that $m = 2^{q-1} + 1$. Moreover, let

$$t = \sigma(\#, u_1\#, x) \circ \sigma(v_1\#, u_2\#, x) \circ \cdots \circ \sigma(v_{m-2}\#, u_{m-1}\#, x) \circ \sigma(v_{m-1}\#, \#, \#).$$

Observe that t contains m occurrences of σ , and that $\iota(t) = U_q$. Moreover, the chains of t are of the form $\alpha_1 \cdots \alpha_\ell$, resp. $\beta_1 \cdots \beta_\ell$, for some $\ell \in [q]$.

Proposition 7.1. $t \in L$.

Proof. Algorithm 1 demonstrates how to derive $h(t)$ in G_{ex} . There, $\text{apply}(\pi, \eta)$ denotes the parallel application of the production π in P_{ex} to every possible position in the sentential form η . Clearly, the result is $\eta = h(t)$, and therefore $t \in L$. \square

As $m > 1$, there are $t_1, t_2 \in T(\Sigma)_1^1$ such that

$$A_\varepsilon \cdot \chi \Rightarrow_G^* B_\varepsilon \cdot s_\varepsilon \cdot \chi \Rightarrow_G (A_1 \cdot \vartheta_1 \cdot s_\varepsilon \circ A_2 \cdot \vartheta_2 \cdot s_\varepsilon) \cdot \chi \Rightarrow_G^* t_1 \circ t_2 = t.$$

Since both t_1 and t_2 contain at least one occurrence of σ , there is a $j \in [m-1]$ such that

$$\begin{aligned} t_1 &= \sigma(\#, u_1\#, x) \circ \sigma(v_1\#, u_2\#, x) \circ \cdots \circ \sigma(v_{j-1}\#, u_j\#, x) \quad \text{and} \\ t_2 &= \sigma(v_j\#, u_{j+1}\#, x) \circ \cdots \circ \sigma(v_{m-2}\#, u_{m-1}\#, x) \circ \sigma(v_{m-1}\#, \#, \#), \end{aligned}$$

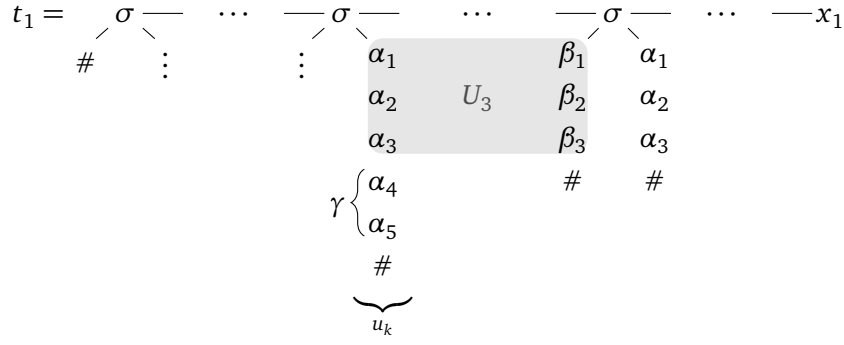


Figure 1: Occurrence of a defect γ in the critical chain u_k of t_1

and this factorization of t induces an according factorization of Z_q into Z' and Z'' with

$$Z' = \langle u_1, v_1, \dots, u_j \rangle \quad \text{and} \quad Z'' = \langle v_j, \dots, u_{m-1}, v_{m-1} \rangle.$$

Example 7.1. Let us consider an example which relates the introduced concepts. Figure 1 displays the critical chain u_k in t_1 , whose defect is $\gamma = \alpha_4 \alpha_5$. As u_k is critical, every a -chain $u_{k'}$ in t_1 to its right (i.e., with $k' > k$) is of the form $\alpha_1 \cdots \alpha_\ell$, for some $\ell \leq 3$. In our intuition, γ is a sequence of opening parentheses which have no corresponding closing parenthesis in t_1 . Therefore, t_2 must contain a suitable sequence of closing parentheses. Formally, γ^R occurs in $P_{i,j}$ as

$$P_{i,j} = P' U_5 \alpha_5 \alpha_4 U_3 P'', \quad \text{so} \quad D_{i,j} = D' \$ \gamma^R \$ D'',$$

for some $P', P'' \in \Gamma^*$ and $D', D'' \in (\Gamma \cup \{\$\})^*$. Therefore, γ is indeed a defect by definition.

By Proposition 6.2(2), the number of defects in $D_{q,j}$ is $q + 1 = 2p + 1$. Thus either t_1 contains at least $p + 1$ critical chains, or t_2 does.

For the rest of this work, assume that t_1 contains at least $p + 1$ critical chains. The proofs for the other case are obtained mainly by substituting b for a and β for α .

By Proposition 5.1, there are a $\hat{t} \in T(\Sigma)_{p+1}^1$ with $t = \hat{t} \cdot \chi$, and a derivation tree κ of \hat{t} . Note that the height of κ is at most m . Therefore $|\delta| < m$ for every $\delta \in \text{pos}(\kappa)$. If $\delta = i_1 \cdots i_d$, then we denote the prefix $i_1 \cdots i_{d-\ell}$ of δ by δ_ℓ , for every $\ell \in [0, d]$. In particular, $\delta_0 = \delta$ and $\delta_d = \varepsilon$.

Let $s \in S(\Gamma)_p^p$ and $w \in \Gamma^*$. If there is no possibility of confusion, we will briefly say that w is a component of s if s has a component of the form wx_i , for some $i \in [p]$.

Proposition 7.2. *Let u_i be an a -chain of t_1 , with $i \in [j]$. There is a leaf δ of κ such that*

$$u_i = w_0 \cdots w_d,$$

where $d = |\delta|$, and w_ℓ is a component of s_{δ_ℓ} , for $\ell \in [0, d]$. Moreover, $\delta_{d-1} = 1$.

Proof. Every leaf node of κ contributes exactly one occurrence of σ to t . So the chain u_i is contributed to t by κ 's i -th leaf node δ , when enumerated from left to right. Let $d = |\delta|$. By tracing the path from δ to the root of κ , we see that

$$u_i\# = \pi_{j_{\delta_0}} \cdot s_{\delta_0} \cdot \vartheta_{\delta_0} \cdots s_{\delta_{d-1}} \cdot \vartheta_{\delta_{d-1}} \cdot s_\varepsilon \cdot \chi.$$

Therefore $u_i = w_0 \cdots w_d$, where w_ℓ is a component of s_{δ_ℓ} , for each $\ell \in [0, d]$. \square

In particular, w_d is a component of s_ε . The next lemma is a consequence of the fact that s_ε has only p components apart from x_{p+1} .

Lemma 7.1. *There is an a -defect γ whose critical chain is of the form $w'w$ for some $w', w \in \Gamma^*$ such that w is a component of s_ε , and $|\gamma| > |w| + mH$.*

Proof. Since t_1 contains more than p critical chains, by Proposition 7.2 there must be two critical chains, say $u\gamma\alpha_i$ and $u'\gamma'\alpha_j$, where $\gamma\alpha_i$ and $\gamma'\alpha_j$ are distinct a -defects with $i < j$, such that

$$u\gamma\alpha_i = w'w \quad \text{and} \quad u'\gamma'\alpha_j = w''w \quad \text{for some } w', w'' \in \Gamma^*,$$

and some component w of s_ε .

Observe that α_i is not a suffix of w , as otherwise α_i would be a suffix of α_j . Therefore $|w| < |\alpha_i|$, and hence

$$|w| + mH < |\alpha_i| + mH = |\alpha_{i+1}| \leq |\alpha_j| \leq |\gamma'\alpha_j|.$$

So the a -defect $\gamma'\alpha_j$ satisfies the properties in the lemma. \square

Theorem 7.1. *There is some $t' \in \mathcal{L}(G) \setminus L$.*

Proof. Let γ be the a -defect from Lemma 7.1. Assume that γ 's critical chain in t_1 is u_k , where $k \in [j]$. Then $u_k = w_0 \cdots w_d$, where w_ℓ is a component of s_ℓ , for each $\ell \in [0, d]$. Moreover, $|\gamma| > |w_d| + mH$. Let f be the largest number such that $w_f \cdots w_d$ has γ as suffix, then $f \in [0, \dots, d-1]$, and there are $w, w' \in \Gamma^*$ such that $w_f = w'w$ and $\gamma = ww_{f+1} \cdots w_d$.

Since $d < m$ and $|ww_{f+1} \cdots w_{d-1}| > mH$, there is a $\tilde{w} \in \{w, w_{f+1}, \dots, w_{d-1}\}$ such that $|\tilde{w}| > H$. In other words, there is an $\ell \in [f, d-1]$ such that $A_{\delta_\ell} \Rightarrow_G^* B_{\delta_\ell} \cdot s_{\delta_\ell}$, and there is some $i \in [p]$ such that either (i) $\ell = f$ and $\pi_i \cdot s_{\delta_\ell} = w'\tilde{w}x_i$, or (ii) $\ell \neq f$ and $\pi_i \cdot s_{\delta_\ell} = \tilde{w}x_i$. In both cases Lemma 5.1 can be applied, and we receive that $s_{\delta_\ell} = v \cdot y \cdot z$, and by pumping zero times, also $A_{\delta_\ell} \Rightarrow_G^* B_{\delta_\ell} \cdot v \cdot z$. Therefore a derivation tree κ' can be constructed from κ by replacing the tuple s_{δ_ℓ} by $v \cdot z$. As δ_ℓ begins with the symbol 1, this alteration does only concern t_1 , thus κ' derives a tree $\tilde{t}' \in T(\Sigma)_{p+1}^1$ such that $\tilde{t}' \cdot \chi = t'_1 \circ - t_2$, for some $t'_1 \in T(\Sigma)_1^1$. Denote $\tilde{t}' \cdot \chi$ by t' .

Let us compare the k -th a -chain u'_k of t'_1 to u_k . Assume that the i -th components of v , y , and z are, respectively, $v'x_i$, $y'x_i$ and $z'x_i$. Then in case (i), there is a $w'' \in \Gamma^*$ such that $v' = w'w''$, as $y'z'$ is a suffix of w . Therefore,

$$u_k = w_1 \cdots w' \underbrace{w'' y' z' w_{f+1} \cdots w_d}_\gamma \quad \text{and} \quad u'_k = w_1 \cdots w' w'' z' w_{f+1} \cdots w_d.$$

In case (ii),

$$u_k = w_1 \cdots w_{f+1} \underbrace{w_{f+1} \cdots w_{\ell-1} v' y' z' w_{\ell+1} \cdots w_d}_{\gamma} \quad \text{and} \quad u'_k = w_1 \cdots w_{\ell-1} v' z' w_{\ell+1} \cdots w_d.$$

It is easy to see that $|t'|_{\sigma} = |t|_{\sigma}$, as the shape of κ was not modified. Thus Proposition 3.2 implies that if $t' \in L$, then also $|t'|_c = |t|_c$ and $|t'|_d = |t|_d$. In particular, $y' \in a^*$. Therefore, both in case (i) and (ii), $P'_{i,j} = \iota(t'_1)$ is a perturbation of $P_{i,j}$. Say that $P_{i,j}$ and $P'_{i,j}$ are of the form as in (4). Since $|y'| > 0$ by Lemma 5.1, at least one a was removed from the occurrence of γ^R in $P_{i,j}$. Therefore, there is some $e \in [q]$ such that $\alpha_e \neq \alpha'_e$. By Lemma 6.1, therefore $P'_{i,j} \not\equiv P_{i,j}$, and hence

$$\iota(t') \equiv \iota(t'_1) \iota(t_2) \not\equiv \iota(t_1) \iota(t_2) \equiv \varepsilon.$$

So $\iota(t') \notin D_{\Gamma}^*$, and by Proposition 3.1, $t' \notin L$. □

Therefore, there is no cftg G with $\mathcal{L}(G) = h^{-1}(\mathcal{L}(G_{\text{ex}}))$, and we have proven Theorem 3.1.

8 Linear monadic context-free tree languages and inverse homomorphisms

In this section, we close the paper with the positive result announced in the introduction.

Theorem 8.1. *The class of linear monadic context-free tree languages is closed under inverse linear tree homomorphisms.*

We will prove this theorem in the remainder of this section. As the constructions are not very difficult, we use a style that is not so formal. Let us start out with recalling a normal form for lm-cftg given in [9].

Let $G = (N, \Delta, \eta_0, P)$ be an lm-cftg.⁶ We say that G is in *Greibach normal form* if $\eta_0 = S$ for some $S \in N^{(0)}$ and each production in P is of one of the following forms:

$$(G1) \quad A \rightarrow \alpha \text{ for some } A \in N^{(0)}, \alpha \in \Delta^{(0)},$$

$$(G2) \quad A \rightarrow \delta(B_1, \dots, B_{i-1}, \eta, B_{i+1}, \dots, B_k) \text{ for some } A \in N^{(0)}, \text{ and } \eta \in T(N)_0^1, \text{ or}$$

$$(G3) \quad A(x) \rightarrow \delta(B_1, \dots, B_{i-1}, \eta, B_{i+1}, \dots, B_k) \text{ for some } A \in N^{(1)} \text{ and } \eta \in \tilde{T}(N)_1^1,$$

and $k \in \mathbb{N}$, $i \in [k]$, $\delta \in \Delta^{(k)}$, $B_1, \dots, B_{i-1}, B_{i+1}, \dots, B_k \in N^{(0)}$. Note that every Greibach cftg is nondeleting.

Lemma 8.1 ([9, Theorem 4.3]). *For every lm-cftg G there is an equivalent lm-cftg G' in Greibach normal form.*

The following decomposition theorem admits proving Theorem 8.1 in a modular manner.

⁶Where Σ and Δ will denote arbitrary ranked alphabets in the following, unless stated otherwise.

Lemma 8.2 ([5, Lemma 10]). *Let $h: T_\Sigma(X) \rightarrow T_\Delta(X)$ be a linear tree homomorphism. There are a linear alphabetic tree homomorphism φ , as well as elementary ordered tree homomorphisms ψ_1, \dots, ψ_k for some $k \in \mathbb{N}$ such that $h = \psi_k \circ \dots \circ \psi_1 \circ \varphi$.*

So in order to show that the linear monadic context-free tree languages are closed under inverse linear tree homomorphisms, it suffices to show closure under these two restricted types.

Lemma 8.3. *The class of linear monadic context-free tree languages is closed under inverse linear alphabetic tree homomorphisms.*

Proof. Assume an lm-cftg $G = (N, \Delta, S, P)$ in Greibach normal form, and let $h: T_\Sigma(X) \rightarrow T_\Delta(X)$ be a linear alphabetic tree homomorphism. Let $H = (M, \Sigma, Z, R)$ be a regular tree grammar such that $\mathcal{L}(R) = T_\Sigma$, and M is disjoint from N . We use the same idea as in [2, Theorem 4.1] to construct an lm-cftg $G' = (N', \Sigma, S, P')$ with $\mathcal{L}(G') = h^{-1}(\mathcal{L}(G))$. Let $N' = N \cup M \cup \{E^{(1)}\}$ such that $E \notin N \cup M$, and let P' be given as follows.

- (i) For every production of type (G1) in P , P' contains $A \rightarrow E(\alpha)$.
- (ii) For every production of type (G2) in P , if $h(\sigma) = \delta(x_{j_1}, \dots, x_{j_k})$ for some $n \in \mathbb{N}$, $\sigma \in \Sigma^{(n)}$, and $j_1, \dots, j_k \in [n]$, then P' contains $A \rightarrow E(\sigma(u_1, \dots, u_n))$, where for each $\ell \in [n]$,

$$u_\ell = \begin{cases} B_m & \text{if } \ell = j_m \text{ for some } m \neq i, \\ \eta & \text{if } \ell = j_i, \\ Z & \text{if } \ell \notin \{j_1, \dots, j_n\}. \end{cases}$$

- (iii) The analogous applies to every production of type (G3).
- (iv) For every $n \in \mathbb{N}$ and $\sigma \in \Sigma^{(n)}$ such that $h(\sigma) = x_j$ for some $j \in [n]$, P' contains the production $E(x) \rightarrow \sigma(u_1, \dots, u_n)$, where for each $\ell \in [n]$,

$$u_\ell = \begin{cases} x & \text{if } \ell = j, \\ Z & \text{otherwise.} \end{cases}$$

- (v) P' contains the productions $E(x) \rightarrow x$ and $E(x) \rightarrow E(E(x))$.

The equivalence of G and G' is shown by defining a tree homomorphism $\varphi: T_{N \cup \{E\}}(X) \rightarrow T_N(X)$ such that $\varphi(E) = x$, and $\varphi(A) = A$ for every $A \in N$. Then it is easy to prove by induction on the length on the derivations that for every $\eta \in T(N \cup \{E\})_0^1$ and $t \in T(\Sigma)_0^1$ we have $\eta \Rightarrow_{G'}^* t$ if and only if $\varphi(\eta) \Rightarrow_G^* h(t)$. \square

Lemma 8.4. *The class of linear monadic context-free tree languages is closed under inverse elementary ordered tree homomorphisms.*

Proof. For this purpose, let Ω be a ranked alphabet such that Ω and $\{\delta_1, \delta_2, \sigma\}$ are disjoint. Let $\Sigma = \Omega \cup \{\sigma^{(k)}\}$ and $\Delta = \Omega \cup \{\delta_1^{(n-k+1)}, \delta_2^{(k)}\}$ for some $n, k \in \mathbb{N}$. Let $h: T_\Sigma(X) \rightarrow T_\Delta(X)$ be the elementary ordered tree homomorphism with

$$h(\sigma(x_1, \dots, x_n)) = \delta_1(x_1, \dots, x_{\ell-1}, \delta_2(x_\ell, \dots, x_{\ell+k-1}), x_{\ell+k}, \dots, x_n)$$

for some $\ell \in [n+1]$, and h is the identity on Ω .

Assume an lm-cftg $G = (N, \Delta, S, P)$. We will construct an lm-cftg G' such that $\mathcal{L}(G') = h^{-1}(\mathcal{L}(G))$. We proceed in several steps.

First, we construct an lm-cftg G_3 with $\mathcal{L}(G_3) = \mathcal{L}(G)$ that fulfills the following property (P): in the right-hand side of each production of G_3 the terminals δ_1 and δ_2 either occur together, directly beneath each other, or none of these terminals occurs. Formally, we demand for every production $u \rightarrow v$ of G_3 that $v(\varepsilon) \neq \delta_2$ and for every $w \in \text{pos}(v)$,

$$v(w) = \delta_1 \quad \text{iff} \quad w\ell \in \text{pos}(v) \quad \text{and} \quad v(w\ell) = \delta_2.$$

We assume that G is in Greibach normal form. Moreover, we can assume without loss of generality that

- $\mathcal{L}(G) \subseteq h(T_\Sigma)$,⁷
- that G is total (by Lemma 2.2), and
- that G has no unreachable nonterminal symbols, i.e. for every $A \in N$, there are $\eta \in \tilde{T}(N \cup \Delta)_1^1$ and $\kappa \in T(N \cup \Delta)$ such that $S \Rightarrow_G^* \eta \cdot A \cdot \kappa$ (cf. [5, Proposition 14]).

Then, the following property holds for G .

Observation 8.1 (cf. [2, Lemma 17]). *For all $A \in N$ and $t \in T(\Delta \cup N)_1^1$ with $A \Rightarrow_G^* t$ we have that t has no subtree of one of the following shapes:*

- $\gamma \cdot [u, \delta_2 \cdot v, w]$ for some $\gamma \in \Delta \setminus \{\delta_1\}$
- $\delta_1 \cdot [u, \gamma \cdot v, w]$ for some $\gamma \in \Delta \setminus \{\delta_2\}$ such that $u \in T(\Delta \cup N)_1^{\ell-1}$, or
- $\delta_1 \cdot [u, \delta_2 \cdot v, w]$ with $u \in T(\Delta \cup N)_1^m$ and $m \neq \ell - 1$,

and where $u, v, w \in T(\Delta \cup N)$.

Let in the following

$$\tilde{N} = \{A \in N \mid \exists u \in T(\Sigma)_k^1: A \Rightarrow_G^* \delta_2 \cdot u\}.$$

As G is total and $\mathcal{L}(G) \subseteq h(T_\Sigma)$, the following observation can be made.

Observation 8.2. *Let $A \in \tilde{N}$. For every $t \in \mathcal{L}(G, A)$, we have $t(\varepsilon) = \delta_2$.*

⁷If it is not, one can apply a similar method to the one in [20, Theorem 2] to construct a cftg G' with $\mathcal{L}(G') = \mathcal{L}(G) \cap h(T_\Sigma)$. Note that G' is again linear and monadic.

We now construct the lm-cftg $G_1 = (N_1, \Sigma, S, P_1)$ with $N_1 = N \cup \{C_\rho \mid \rho \in P\}$ and the following productions in P_1 :

- (i) Every production $A \rightarrow t$ in P with $t(\varepsilon) \neq \delta_2$ is also in P_1 .
- (ii) For every production $\rho = A \rightarrow \delta_2(B_1, \dots, B_{i-1}, \eta, B_{i+1}, \dots, B_k)$ in P , with $\eta \in T(N)_0^1$, the productions $A \rightarrow \delta_2(B_1, \dots, B_{i-1}, C_\rho, B_{i+1}, \dots, B_k)$ and $C_\rho \rightarrow \eta$ are in P_1 .
- (iii) For every production $\rho = A(x) \rightarrow \delta_2(B_1, \dots, B_{i-1}, \eta, B_{i+1}, \dots, B_k)$ in P , with $\eta \in \tilde{T}(N)_1^1$, the productions $A(x) \rightarrow \delta_2(B_1, \dots, B_{i-1}, C_\rho(x), B_{i+1}, \dots, B_k)$ and $C_\rho(x) \rightarrow \eta$ are in P_1 .

It is easy to see that $\mathcal{L}(G_1) = \mathcal{L}(G)$. Now consider a production

$$\rho' = A \rightarrow \delta_2(B_1, \dots, B_{i-1}, C_\rho, B_{i+1}, \dots, B_k)$$

in P_1 . By Observations 8.1 and 8.2, $B_j \neq A$ for each $j \in [k] \setminus \{i\}$. For this reason and since C_ρ is a fresh nonterminal, A does not occur in the right-hand side of ρ' .

Thus, we can *eliminate* the production ρ' from G_1 , as described in [18, Def. 11]. We construct an lm-cftg $\text{Elim}(G_1, \rho')$ as follows: for each production $s \rightarrow t$ in P_1 and each $W \subseteq \{w \in \text{pos}(t) \mid t(w) = A\}$, we construct a new production $s \rightarrow t'$ and insert it into P_1 . The new right-hand side t' is obtained by substituting the right-hand side of ρ' for A at each position in W . Then ρ' is removed from P_1 . It was shown in [18, Lemma 12] that $\mathcal{L}(\text{Elim}(G_1, \rho')) = \mathcal{L}(G_1)$. The same idea works for productions of the form $A(x) \rightarrow \delta_2(B_1, \dots, B_{i-1}, C_\rho(x), B_{i+1}, \dots, B_k)$ in P_1 .

As an example, when we eliminate the production $\rho' = C(x) \rightarrow \delta_2(A, B(x))$ in G_1 , the production $A(x) \rightarrow \delta_1(B, C(D(x)), E)$ in G_1 results in *two* new productions,

$$\rho_1 = A(x) \rightarrow \delta_1(B, C(D(x)), E) \quad \text{and} \quad \rho_2 = A(x) \rightarrow \delta_1(B, \delta_2(A, B(D(x))), E),$$

and ρ' is discarded.

By applying this procedure successively for each production with a nonterminal from \tilde{N} on its left-hand side, we obtain in finitely many steps an equivalent lm-cftg G_2 . Note that G_2 “nearly” has property (P): it may still contain productions which are not of the desired form.

In our example, if ρ' was the last production to be eliminated, then there is still the production ρ_1 left, where δ_2 does not occur under δ_1 . However, it is easy to see that this production is *useless*: after all, there are no productions left for the nonterminal C .

This observation applies to all productions $s \rightarrow t$ which are not of the desired form. Therefore, $\mathcal{L}(G, t) = \emptyset$, and by a lemma of Rounds [24, p. 113], we can just remove all these useless productions, resulting in the lm-cftg $G_3 = (N_3, \Delta, S, P_3)$, which has property (P).

* * *

We now use the same idea as in [5]. As δ_1 and δ_2 appear right beneath each other in the productions of G_3 , they can just be replaced by σ .

Formally, define a homomorphism $\varphi: T_{N_3 \cup \Sigma}(X) \rightarrow T_{N_3 \cup \Delta}(X)$ such that $\varphi(A) = A$ for each $A \in N_3$ and $\varphi|_\Sigma = h$. We construct an lm-cftg $G' = (N_3, \Sigma, S, P')$ such that for each $A \in N_3$ and $t \in T(N_3 \cup \Sigma)_1^1$, we have that $A \rightarrow t$ is in P' if and only if $A \rightarrow \varphi(t)$ is in P_3 .

The formal proof that $\mathcal{L}(G') = h^{-1}(\mathcal{L}(G))$ is omitted. □

9 Conclusion

In this work, we proved that the class of linear context-free tree languages is not closed under inverse linear tree homomorphisms. However, the tree languages of linear monadic context-free tree grammars, which are employed in praxis under the pseudonym of tree-adjoining grammars, are closed under this operation.

In applications which require nonmonadicity and closure under inverse homomorphisms, it may prove beneficial to revisit the formalism of *k-algebraic grammars*, i.e. context-free tree grammars over magmoids, where a nonterminal may derive a tuple of trees [3, Chapter V]. The class of languages defined by this type of grammar is indeed closed under inverse linear tree homomorphisms.

Acknowledgement Last but foremost, we want to thank André Arnold for his help in this work. In our email conversations, which we enjoyed very much, he showed us the flaws in our first proof attempts, and encouraged us to try on. Moreover, the idea for the intermediate normal form of G in Lemma 4.2 is due to him, and he showed us how to significantly improve the presentation of the results in Sections 6 and 7.

References

- [1] A. Arnold and M. Dauchet. Translations de Forêts Reconnaisables Monadiques; Forêts Corégulières. *RAIRO – Informatique Théorique*, 10:5–21, 1976.
- [2] A. Arnold and M. Dauchet. Forêts Algébriques et Homomorphismes Inverses. *Information and Control*, 37:182–196, 1978.
- [3] A. Arnold and M. Dauchet. Théorie des Magmoïdes. *RAIRO – Theoretical Informatics and Applications – Informatique Théorique et Applications*, 12(3):235–257, 1978, and 13(2):135–154, 1979.
- [4] A. Arnold and M. Dauchet. Morphismes et Bimorphismes d’Arbres. *Theoretical Computer Science*, 20:33–93, 1982.
- [5] A. Arnold and B. Leguy. Une Propriété des Forêts Algébriques «de Greibach». *Information and Control*, 46(2):108–134, 1980.
- [6] A. Arnold and M. Nivat. Formal Computations of Non Deterministic Recursive Program Schemes. *Mathematical Systems Theory*, 13(1):219–236, 1979.
- [7] J. Engelfriet and E. M. Schmidt. IO and OI. *Journal of Computer and System Sciences*, 15(3):328–353, 1977, and 16(1):67–99, 1978.
- [8] M. J. Fischer. *Grammars with Macro-Like Productions*. PhD thesis, Harvard University, 1968.

- [9] A. Fujiyoshi. Analogical Conception of Chomsky Normal Form and Greibach Normal Form for Linear, Monadic Context-Free Tree Grammars. *IEICE – Transactions on Information and Systems*, E89-D(12):2933–2938, 2006.
- [10] Z. Fülöp, A. Maletti, and H. Vogler. Weighted Extended Tree Transducers. *Fundamenta Informaticae*, 111:163–202, 2011.
- [11] K. Gebhardt and J. Osterholzer. A Direct Link between Tree Adjoining and Context-Free Tree Grammars. In *Proceedings of the 12th International Conference on Finite-State Methods and Natural Language Processing*, 2015.
- [12] J. A. Goguen, J. W. Thatcher, E. G. Wagner, and J. B. Wright. Initial Algebra Semantics and Continuous Algebras. *Journal of the ACM*, 24(1):68–95, 1977.
- [13] M. A. Harrison. *Introduction to Formal Language Theory*. Addison-Wesley, Boston, MA, USA, 1978.
- [14] M. Kanazawa. Multi-Dimensional Trees and a Chomsky-Schützenberger-Weir Representation Theorem for Simple Context-Free Tree Grammars. Technical Report 3, National Institute of Informatics, Japan, 2013.
- [15] S. Kepser and U. Mönnich. Closure Properties of Linear Context-Free Tree Languages with an Application to Optimality Theory. *Theoretical Computer Science*, 354(1):82–97, 2006.
- [16] S. Kepser and J. Rogers. The Equivalence of Tree Adjoining Grammars and Monadic Linear Context-Free Tree Grammars. *Journal of Logic, Language and Information*, 20(3):361–384, 2011.
- [17] T. S. E. Maibaum. A Generalized Approach to Formal Languages. *Journal of Computer and System Sciences*, 8(3):409–439, 1974.
- [18] A. Maletti and J. Engelfriet. Strong Lexicalization of Tree Adjoining Grammars. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 506–515, 2012.
- [19] A. Maletti, J. Graehl, M. Hopkins, and K. Knight. The Power of Extended Top-Down Tree Transducers. *SIAM Journal on Computing*, 39(2):410–430, 2009.
- [20] M.-J. Nederhof and H. Vogler. Synchronous Context-Free Tree Grammars. In *Proceedings of the 11th International Workshop on Tree Adjoining Grammars and Related Formalisms*, pages 55–63, 2012.
- [21] M. Nivat. On the Interpretation of Recursive Program Schemes. Notes of a lecture given at the Advanced Course on Semantics of Programming Languages, Saarbrücken, 1974.

- [22] J. Osterholzer. Complexity of Uniform Membership of Context-Free Tree Grammars. In A. Maletti, editor, *Algebraic Informatics*, volume 9270 of *Lecture Notes in Computer Science*, pages 176–188. Springer, 2015.
- [23] W. C. Rounds. Mappings and Grammars on Trees. *Theory of Computing Systems*, 4(3):257–287, 1970.
- [24] W. C. Rounds. Tree-Oriented Proofs of Some Theorems on Context-Free and Indexed Languages. In *Proceedings of the Second Annual ACM Symposium on Theory of Computing*, pages 109–116, 1970.
- [25] W. C. Rounds. Complexity of Recognition in Intermediate Level Languages. In *Proceedings of the 14th Annual Symposium on Switching and Automata Theory*, 1973.